

Better climate data required

The global observing system provides the raw data climate research relies on. But the quality of the gathered data is patchy. The system needs improvement — on land and on sea.

In January, the Intergovernmental Panel on Climate Change (IPCC) startled the world with a frightening climate scenario: by 2100, global temperatures could rise by almost 6 °C. Fortunately, this is only a worst-case scenario. Other models predict a temperature rise of only 1.4 °C. Most probably, the rise in the next decade will lie between these extremes. But there is also a small chance that none of the IPCC's scenarios will come close to reality.

So is climate prediction no more than a game of chance? Not quite — but it is closer to one than it need be. The accuracy of any model depends significantly on the quality of the underlying raw data. The problem is, the quality is patchy.

Monitoring and predicting the global climate requires a reliable system of constant observations, rapid data exchange and long-term recording, in standardized formats. That is why, in 1992, the Global Climate Observing System (GCOS) was established. Its aim is to ensure that climate-relevant information is obtained and made available to all potential users. The creation of GCOS was a major advance, but the reporting system has significant deficiencies, and there are large gaps in global and regional coverage, which seriously affect climate assessment and modelling efforts.

Terrestrial climate monitoring is currently based on a network of around 1,000 GCOS observation stations. But the reliability of the data that some of them collect is inadequate, and a disproportionate number of these stations are in rich countries, with sparse coverage in many regions of Africa, South America and Asia. Although no one can say how many climate data are needed for accurate monitoring, clearly the geographical distribution should be as even as possible.

But a GCOS station costs up to \$500,000 per year to operate and maintain — too much for poorer countries. And maintenance costs are particularly high in remote polar regions; over the past years, several

stations have been closed, for example, in Russia and Canada.

Under the United Nations Framework Convention on Climate Change, all countries are required to set up and run appropriate observation programmes, and to exchange data with other nations and with international organizations. But in practice, many poorer countries spend little on their regional climate observations. The training of technical staff and the maintenance of instruments at their observation stations are often inadequate. Through misreporting, instrumental drifts and biases, unreliable communication infrastructures or political unrest, about half of the world's climate data potential is lost or corrupted each month.

The World Meteorological Organization gives regional training and technical support. But this is not enough. The United Nations should also help poor countries to collect and distribute accurate and consistent data sets.

Sea-based climate observation and ocean monitoring, which is likely to add significantly to our knowledge of what drives atmospheric processes, is only just beginning. The deep oceans in particular are still under-observed. Efforts must continue to implement a more systematic ocean observation system. The Global Ocean Observing System (GOOS), founded with GCOS, has the right approach. GOOS coordinates the use of new technologies, such as meteorological buoys, which measure climate-relevant variables in and over the oceans. But it must be expanded.

A worldwide network of sea-based observation buoys will not come cheaply, and will need strong international coordination. Projects such as the Tropical Ocean Global Atmosphere programme, which ended in 1994, have shown that systematic ocean observation is essential for predicting El Niños or seasonal weather. There is a strong case for heavy investment in ocean meteorology. ■

A danger to society

Research into the biological basis of psychopathy deserves higher priority.

Considering the acres of newsprint devoted daily to the exploits of cold-blooded killers, our ignorance of their underlying personality disorder is rather shocking. More than a quarter of the inmates of many high-security prisons can be classified as psychopaths — individuals with an emotional deficit that renders them fearless and lacking in empathy. Many psychiatrists now accept that the underlying condition is biological, and can even be viewed as a 'disease'.

Unfortunately, our understanding of the biological basis of psychopathy remains rudimentary. There has been but a smattering of research into its physiological correlates, and attempts to apply the techniques of neuroimaging to study the brains of psychopaths have so far generated as much heat as light (see page 296). But, at present, we do not even know the incidence of psychopathy in the general population. Could it be true, as some psychopathy experts have only half-jokingly hypothesized, that the traits of psychopathy are actually an advantage in some careers, such as politics?

A comprehensive programme of research — ranging from neuropharmacology, through brain imaging, to traditional psychiatric and behavioural approaches — could help to identify what makes criminal psychopaths notorious recidivists, and how best to treat them.

Such arguments lack popular appeal. The public seems more concerned about punishing violent criminals than rehabilitating them. The idea of sending such people for treatment in secure mental hospitals, rather than to prison, is widely disparaged as a 'soft' option.

We can argue about whether criminal psychopaths are mad, or bad. But there is no doubt that they can be extremely dangerous to know. And the current system, in which violent criminals may be released without having addressed the personality disorder that predisposed them to offend in the first place, serves no one well. There must be a better way, and research into the biological roots of psychopathy may lead us there. ■





Moving story
Natural history
archive heads for
a new home
p290



Added values
Supporters rally
behind the standard
model of physics
p291



Down under
US physicists want
to mine a rich seam
of research
p292



Fears of a clone
Rome-based
doctor plans
human cloning
p293

Free access to cDNA provides impetus for gene function work

Alison Abbott, Munich

Hot on the heels of the publication of the draft human genome sequence, several efforts are under way to provide all scientists with free, synthetic versions of human genes. Their availability will speed up attempts to understand the function of human genes and their associated proteins.

Each of the synthetic genes, called complementary DNAs or cDNAs, can be induced to synthesize the protein encoded from the real gene from which it was derived.

A German team of researchers, half-way through a three-year pilot study, announces in this month's *Genome Research* (11, 422–435) that it has generated 500 previously unknown full-length human cDNAs, and an additional 1,000 known cDNAs. All the cDNAs have been inserted into transportable packages known as vector clones, and are freely available to academic and industrial scientists.

Since these results were submitted, the group, led by Stefan Wiemann of the German Cancer Research Centre in Heidelberg, has

made available over a hundred additional cDNA clones. "Requests for the clones from scientists around the world have increased dramatically this year," says Wiemann.

A typical request, he explains, might come from a cancer researcher who sees very low expression of a particular gene in one type of tumour, and wants to know what it does. If the cDNA for that gene is available as a clone, the researcher can get to work on its function straight away.

The cDNAs are synthesized from 'messenger RNAs'. Each mRNA is a transcript of a gene, and its job is to direct the stringing together of the correct amino acids that constitute the protein encoded by the gene. The enzyme reverse transcriptase is used to direct the synthesis, or reverse transcription, of cDNAs from mRNAs.

The creation of full-length cDNAs, including the entire protein-coding region as well as regulatory regions, is no mean achievement. The enzyme is inefficient *in vitro* — it tends to 'fall off' its substrate before the end of the operation, leaving incomplete cDNAs.



In sequence: cloning complementary DNAs can expedite studies of gene function.

Cloning is also inefficient for long cDNA inserts, so it is easier to make short full-length cDNAs than long ones. The average size of the German cDNAs was about 2.5 kilobases (kb), equivalent to a maximum of 500 amino-acid residues.

The work to systematically produce cDNAs was pioneered by the Kazusa cDNA project in Japan, which began in the mid-1990s. This project selects only long cDNAs to synthesize and clone — they are at least 4 kb, equivalent to proteins larger than about 1,000 amino-acid residues. Of the 2,000 cDNAs produced so far, only about half are full-length. "We see large proteins as being the most interesting biologically," says Osamu Ohara, head of the project, which is based at the Kazusa DNA Research Institute in Chiba. The Japanese cDNA clones are freely available to researchers in academia, but not to those in industry.

In the United States, the National Institutes of Health (NIH) started a large cDNA initiative last year, as part of its Cancer Genome Anatomy Project. The NIH project will provide clones freely to all researchers and, like the German project, is creating its library of cDNAs randomly, rather than through size selection. It will be larger than the German project, and has already registered 2,800 full-length sequences with

Anti-AIDS drugs available 'at cost'

David Dickson

As public pressure mounts for AIDS drugs to be made available to poor countries at an affordable cost, a leading drug company has pledged for the first time to sell its protease-inhibitor therapies to them at cost price.

Merck said last week that it will make Crixivan (indinavir sulphate) — a component of the triple-drug anti-AIDS therapy widely used in the West — available at \$600 a year in developing countries. Another antiretroviral agent, Stocrin (efavirenz), is to be sold at \$500. Other major drug companies are expected to follow suit.

The move is an attempt to head off mounting criticism over the cost of AIDS drugs in developing countries (see *Nature* 410, 3; 2001). The announcement came as a court case, brought by 40 drug companies

against the South African government, threatened to turn into a public-relations debacle for the companies.

Merck says that developing countries taking up its offer must introduce procedures to prevent the re-exportation of the drugs to the developed world, where their official price is at least an order of magnitude higher.

"Even at these new lower prices, the issue of access is not resolved," Per Wold-Olsen, president of Merck's human-health business in Europe, the Middle East and Africa, said last week.

Hearings on the legal action, which seeks to block a South African law that would allow the importation of generic copies of patented drugs, opened briefly in South Africa last week, but were adjourned until next month. Kenya is reported to be considering legislation along similar lines.

► GenBank — although some of these may not be unique.

The ultimate goal of all these projects is to produce a complete set of full-length human cDNAs, corresponding to all the genes in the human genome. As there are now thought to be only about 30,000 genes, and technologies needed to deal with long or rare cDNAs are improving, this may not take very long — “perhaps only a couple of years”, suggests an optimistic Robert Strausberg, head of the NIH project.

With this rapid progress, the need to coordinate these efforts has become more pressing. “It would be nice to have a system whereby different sets of cDNAs were allocated to different groups, so there would not be too much duplication of effort,” says Wiemann. Ohara agrees: “A couple of years ago we were the only group in the game, but now we really need to think of an allocation system,” he says.

There could be complications in organizing such a system, Strausberg points out, because not all groups make their results freely available to all. The issue is to be discussed in the next few months by the various projects. In the meantime, a joint website is planned to help each project keep up with what the others are doing. ■

► <http://www.rzpd.de>

► <http://cgap.nci.nih.gov>

► <http://www.kazusa.or.jp/huge>

Network aims to link species data from global collections

Georgina Kenyon, London

A global programme has been launched to collate information on the world's most important collections of animal and plant species, in a format that is freely accessible to all.

The Global Biodiversity Information Facility (GBIF), which hopes to provide unrestricted access on its website to species information held in existing museum collections, was officially launched in Brussels last week.

The GBIF has received US\$2 million of initial support from its twelve member countries. Subscriptions to the GBIF differ from country to country, depending on their wealth.

It is hoped that the facility will enable scientists in developing countries to have access to the best information about their local biodiversity, often for the first time.

“The complexity of this initiative is vast,” says Martin Sharman, a spokesman for the GBIF. About 1.75 million species have been scientifically described in different formats in various collections, but between 10 million and 100 million are thought to exist.

Organizers will decide in July where to

place GBIF's secretariat, with Australia, the Netherlands, Denmark and Spain all vying to host it.

“GBIF will assist museums, scientists and educators, and give politicians scientifically factual data,” says Carlos Martinez-Riera, a representative of the European Commission and GBIF's chief scientific officer.

Eventually the programme aims to set up a web page for each species, with a standard set of information including characteristics, taxonomy and geographical distribution, Martinez-Riera says.

Several national programmes are currently under way to develop such databases — including Species 2000 in Britain, which is not a full member of the GBIF, but is involved in steering the initiative. At the moment there is a large amount of ‘double counting’, as scientists have independently discovered and named the same species.

Each member country of the GBIF plans to nominate a scientific institution to head its contribution to the project, which was initiated at the Organisation for Economic Co-operation and Development's Megascience Forum in 1996.

► <http://www.gbif.org>

London museum puts its animals on public display

David Adam, London

In a six-month exercise that will provide an important fillip for systemic biology, London's Natural History Museum has started moving its collection of over 20 million zoological samples from storage into a new, purpose-built public archive.

The first specimens being transferred to the £27 million (US\$39 million) archive are the museum's ‘spirit collection’ of fish, reptiles and invertebrates stored in alcohol. Many of these were brought back from Charles Darwin's five-year voyage on the *Beagle* in 1831, and others were collected on Captain James Cook's *Endeavour* expedition in 1763.

The collection includes rays, komodo dragons and tuatara lizards, the last survivors of a group previously considered to have become extinct at the same time as dinosaurs. Many of these specimens will be going on public display for the first time.

The transfer will provide an opportunity for the museum to compile an electronic database of its specimens, in line with efforts to build up a global computerized catalogue of the specimens currently sitting



On show: staff at London's Natural History Museum move a 12-foot oarfish to the Darwin Centre.

in jars and display cases across the world (see story above).

But a global computerized catalogue is some way off, museum officials say. “There is a drive to get some conformity but it will be a long time before those kinds of standards are agreed,” says Paul Henderson, director of science at the museum. “That would be just too labour intensive at the moment because too many institutions have their own systems.”

The London museum is making an effort to standardize species names, however, and is trying to make records about where the specimens came from as accurate as possible.

This first phase of the new archive is expected to open in spring 2002. Phase two will then see the ‘dry’ entomology and botany collections moved to the new facility, which will be known as the Darwin Centre. ■

► <http://www.nhm.ac.uk/darwincentre/index.html>

Theorists reject challenge to standard model

David Adam

Reports of the death of the standard model of particle physics are highly exaggerated, a group of theoretical physicists asserted last week. They questioned recent claims about experimentalists' results that appear to violate the 40-year-old standard model of matter and forces.

A blast of media speculation followed last month's announcement that scientists at the Brookhaven National Laboratory in New York state had obtained data that disagreed with the theoretical result predicted by calculations based on the standard model (see *Nature* **410**, 28–29; 2001).

The group used the world's largest superconducting magnet — the muon storage ring — to help to measure the magnetic moment of particles called muons, which shows how the particles behave in a magnetic field. They say their results are different enough from predicted values to infer physics that the standard model cannot explain.

The press release announcing the finding was headed: "Physicists announce possible violation of Standard Model of particle physics", setting a tone that was quickly echoed by the international press.

At first, sceptical physicists pointed out that the difference lacked the statistical significance normally required in particle physics. But now some theorists charge that the Brookhaven group has been comparing its result with the most convenient theoretical value for the muon's magnetic moment. The sceptics say there are other predicted values for it which are "equally legitimate" — and which would be compatible with the new experimental data.

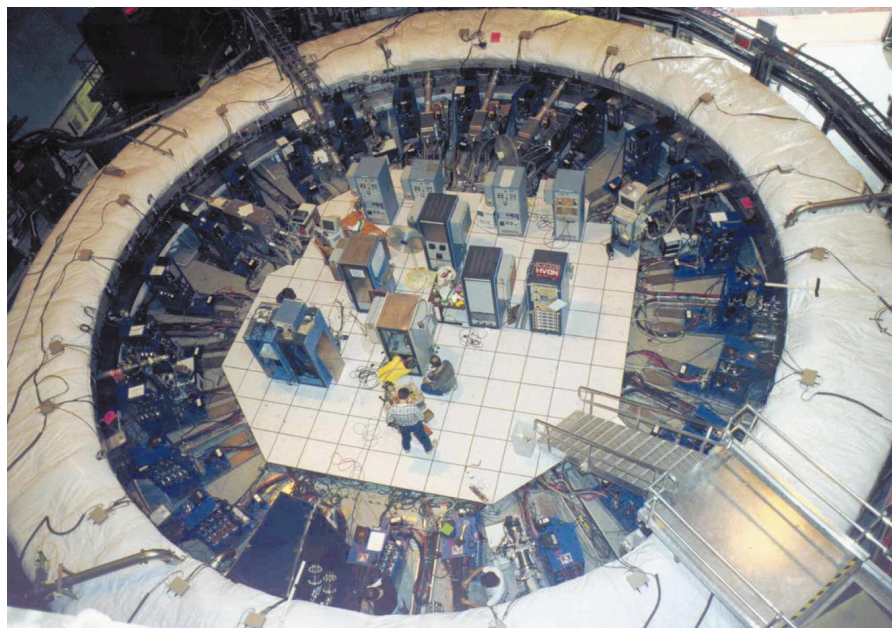
Loaded dice

"These experimentalists are playing with loaded dice," claims Francisco Ynduráin, a theoretical physicist at the Independent University of Madrid. Ynduráin placed a critique of the Brookhaven claim on the Los Alamos preprint server (<http://xxx.lanl.gov/format/hep-ph/0102312>) in February.

"They've chosen the one theoretical calculation that disagrees with their result so as to claim they've discovered something new," Ynduráin says. In fact, he contends, the new experimental data appear to validate the other theoretical values, which broadly agree with one another.

Lee Roberts, a physicist at Boston University and spokesman for the Brookhaven experiment, robustly denies that the team cherry-picked convenient predicted values for comparison.

"The one we used is the latest and what we think is the best value," he says. "I've followed this for over ten years and the [theoretical predictions] have steadily got more precise as



Ring of confidence: but how valuable is the result obtained from Brookhaven's muon experiment?

additional data have been included. I look at the others and I say they are old numbers."

The theoretical values at issue come from a complex set of equations used to predict how rapidly muons wobble in a magnetic field, taking account of the influence of other subatomic particles around them. To do this, theoretical physicists use data both from previous experiments and the mathematical framework of the standard model, which seeks to explain how quarks, neutrinos, electrons, muons and other particles interact.

Roberts says that the accuracy of the theorists' prediction inevitably improves as more experimental data are fed into the equation. But does this mean that the framework underpinning the theorists' work is wrong, if their latest prediction is disproved by an experimental result? Ynduráin, who devised one of the earlier predictions, says not.

"Later is not necessarily better," he says, explaining that each calculation places slightly different emphasis on the components making up the equation and their relative errors. Since he posted his objections, he says that "dozens" of physicists have contacted him to agree.

One of these is Valentine Telegdi, a high-energy physicist at the California Institute of Technology. He says the argument that the most recent theoretical treatment is also the most thorough is "not compelling". And he repeats physicists' initial concerns about the statistical significance of the result.

"The departure from the standard model, even with the theoretical prediction favoured by the Brookhaven group, is statistically weak," says Telegdi. Statistically, the gap between the theoretical and the experi-

mentally obtained results is 2.7 standard deviations, where experimentalists normally like to have three standard deviations before declaring a result. The other theoretical values are about one standard deviation away from the experimental result, which is not statistically significant at all.

Issue not settled

"In my opinion the issue is certainly not settled," says Martinus Veltman, a high-energy physicist based at the University of Michigan, who won the Nobel Prize in Physics in 1979. "The theory and the data used as input for the theory are too complicated to make things as cut and dried as one would want. I tend to agree with Ynduráin that just comparing with the most recent analysis is somewhat too simplistic."

But other physicists support the Brookhaven team's position. They agree that the theoretical prediction based on the most up-to-date information must be the one that best reflects the standard model.

Although accepting that there is "not yet a consensus of opinion", Gordon Kane, a physicist at the University of Michigan, says: "We have to make a choice, and there is a tentative agreement to use [the most recent value] until there is something better."

Kane and Roberts point to new experimental data that should be available within months. These should help both to rework the predicted value and to test the robustness of the Brookhaven results further. If, as Roberts expects, it supports his group's findings, Kane says the implications could be "far more powerful than even optimists are claiming right now". ■

Congress accused of slighting sound science

Mark Schroppe

The law passed by the US Congress in 1996 allowing it to overthrow expensive government regulations was supposed to ensure that any such rules were underpinned by "sound science".

But last week Congress was being accused of ignoring such science, when both houses used the law to throw out rules governing ergonomics in the workplace. Developed by the Occupational Safety and Health Administration (OSHA), the rules were aimed at reducing repetitive-motion injuries by requiring employers to apply the latest ergonomics research.

Supporters of the OSHA regulations claimed that in repealing the rules, Congress was brazenly disregarding the scientific consensus reached in a report on ergonomics, which was published in January by the National Academy of Sciences (NAS).

House minority leader Richard Gephardt (Democrat, Missouri) thinks that the message in the NAS report is clear. "This is a critical issue affecting hundreds of thousands of workers every year," he told the Congress, "and government needs to act in the name of their safety."

The OSHA rules had been under development for a decade before they became effective on 16 January this year. They called for workers to be educated about the dangers



of repetitive-motion tasks such as heavy lifting, operating vibrating equipment and typing. Under certain conditions, they also required employers to alter job specifications so that the tasks fell within established guidelines for injury prevention.

Many businesses and Republicans protested against the rules, saying that they could cost over \$100 billion to implement. The OSHA estimated costs at \$4.5 billion, but claimed that the rules would save companies over \$9 billion through increased productivity.

The Congressionally mandated NAS study, "Musculoskeletal Disorders and the

Workplace: Low Back and Upper Extremities", did not explicitly assess the OSHA rules. But it concluded that repetitive-motion tasks significantly increase risks for certain injuries, and that proper intervention can reduce such risks.

"The whole action of repealing really shows a disregard for the science," says Jonathan Stivers, a spokesman for Representative Nancy Pelosi (Democrat, California), a supporter of the rules. The NAS report "clearly justifies the need" for the new rules, he adds.

Republicans and the handful of Democrats who defeated the rules say that they agree ergonomic problems should be dealt with, but that the rules were poorly designed. "They were just too far-reaching, and were going to be a huge burden on business and industry and cause a lot more problems than they would correct," says Dan Lara, a spokesman for the House Committee on Education and the Workforce.

"My sense of what's happening at the moment is that the [NAS] report is being largely put aside and other agendas are being pursued," says Jeremiah Barondess, president of the New York Academy of Medicine and chair of the panel that produced the NAS report. "The science is there. The science is clear. You can like it or you can not like it, but that's what it is."

Mine mooted as underground home for physicists

Rex Dalton, San Diego

A gold mine in the Black Hills of South Dakota has been selected as the most likely US site for a proposed national underground physics laboratory.

The laboratory, which would cost between \$70 million and \$150 million to build, would be used for highly sensitive experiments in neutrino physics and for other physics experiments that must be conducted away from cosmic rays. Currently, US physicists must travel to the Super-Kamiokande facility at Tsukuba Science City, Japan, or Italy's Gran Sasso Laboratories near Rome to participate in such experiments.

As part of an exploratory effort supported jointly by the National Science Foundation and the Department of Energy, a committee of experts looked at five proposed sites before recommending the Homestake Mine — where the University of Pennsylvania has already conducted some experiments — as the most promising.

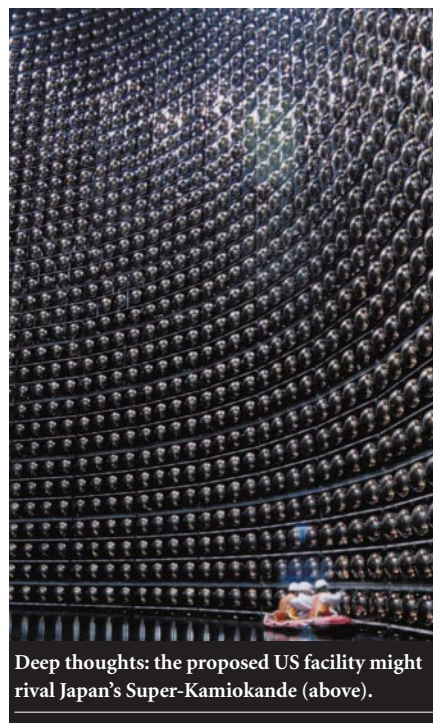
Scheduled to close as a working mine this autumn, the 125-year-old Homestake is economically attractive, as the necessary

access tunnels are already in place. But it also presents certain logistical challenges, because of limitations on the size of equipment that can be transported down its 2,500-metre-deep shafts.

The South Dakota School of Mines and Technology at Rapid City was involved in proposing the Homestake site, both to boost science in the region and as a national resource. But any laboratory that evolves will encompass a broad range of researchers and institutions.

Proponents of the new facility would like to expedite planning so that deep-mine experts can remain on site when the mine closes. But the National Science Foundation and the Department of Energy may struggle to find money for the proposal. The project has yet to be reviewed by agency officials, and state and federal legislation would be needed for the government to assume liability for the mine from its current owners.

Although the panel favoured the South Dakota site, it also recommended environmental assessments of several alternatives in California and Nevada.



Deep thoughts: the proposed US facility might rival Japan's Super-Kamiokande (above).

ICRR/UNIV. TOKYO

Trepidation greets plan for cloning humans

Alison Abbott

Fireworks erupted in Rome last week at an international workshop on human and therapeutic cloning as the meeting's organizers clashed with their critics and the press.

Held at La Sapienza, Rome's largest university, the workshop attracted as many journalists as scientists, and was attended by few of the established experts in cloning technology. Instead, it drew criticism from Italian researchers who feared that it was likely to damage the reputation of Italian biomedical research.

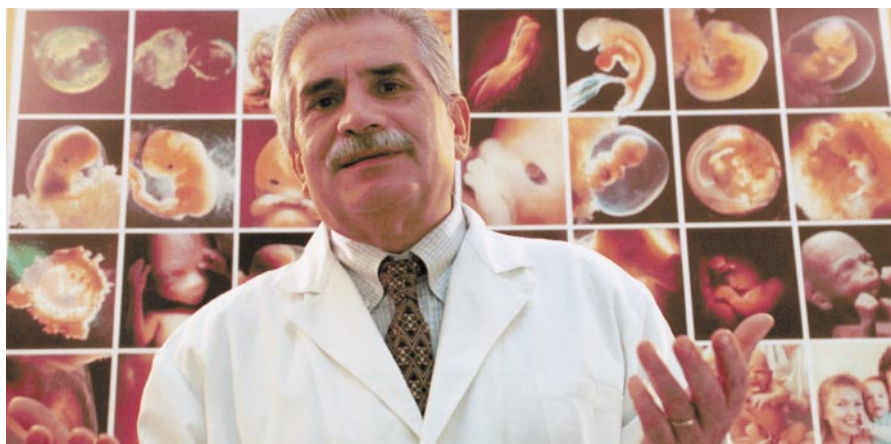
The workshop was organized by Rome-based obstetrician Severino Antinori, Panayiotis Zavos, a fertility researcher from Kentucky, and their new partner, Avi Ben-Abraham, an Israeli physician. The three announced that they would set a date for their first human cloning attempt at a meeting planned for Monaco in October. The experiment is set to take place in an unnamed southern Mediterranean country, they said.

But no details were forthcoming in Rome of the technologies that the trio plans to use to ensure an acceptably high rate of healthy births. Zavos spoke only of reducing errors by using good "quality control" of embryo selection.

Ian Wilmut of the Roslin Institute near Edinburgh, head of the Scottish research team that cloned Dolly the sheep, did not attend the meeting but said afterwards that he doubts that such a reduction is currently feasible. Wilmut noted that low success rates are apparent in each of the five species that have already been cloned, including rhesus monkeys, in which only 4% of reconstructed embryos survive. "We are good at selection," he said, "but our survival rates [in farm animals] are still very low and most pregnancy failures occur just before term, which would be devastating and cruel for humans."

At one point during the meeting, according to several observers, Antinori had to be physically restrained from trying to prevent a young doctor from speaking out against the workshop. The doctor said that a conference on human cloning, which has been barred as unethical by several international organizations and governments, should not have been allowed in a public building, still less in a renowned university such as La Sapienza.

At another point Antinori warned journalists that he was successful in suing those whose reporting he considered defamatory or incorrect. In the week running up to the conference, Antinori brought actions against at least two newspaper writers, including Carlo Redi, a professor at the University of Pavia, and a member of the Italian government advisory committee on stem cells, who writes newspaper columns.



Controversial concept: meeting organizer Severino Antinori is preparing to clone a human being.

Redi believes that the workshop threatens the standing of Italian biomedical science, which, he points out, is already suffering from the so-called Di Bella affair, in which an octogenarian doctor notoriously proclaimed a cure for cancer (see *Nature* 391, 217; 1998).

Redi claims that the announcement of the

workshop in February prompted Umberto Veronesi, the health minister, to ban all cloning work in both humans and animals.

And according to Wilmut, the plan to press ahead with human cloning could contribute to a backlash outside of Italy against important research, such as therapeutic cloning. ■

Singapore invests in bioinformatics

David Cyranoski, Tokyo

Plans for a new institute in Singapore could address one of the most acute skills shortages in science by producing up to 100 trained bioinformaticists a year.

The planned Bioinformatics Institute is part of Singapore's US\$1 billion-a-year effort to turn the island into a powerhouse of biomedical research. Within five years, the institute should be delivering 100 masters degrees in bioinformatics. This is more than any other institution in the world, says Limsoon Wong, director of the Kent Ridge Digital Bioinformatics Laboratories, and one of the planners behind the institute.

According to Wong, training at the institute will go well beyond the curating of data. "We will be training people how to make predictions from the data concerning interaction between proteins, and how to use these data to drive experiments," he says.

The research and teaching institute will be housed temporarily at first, before moving to the planned 'biopolis' science park near the National University of Singapore, when it opens in two years time. The government has yet to announce its funding level, but the park is expected to start with a grant of around S\$100 million (US\$60 million).

The institute is likely to absorb the



Gunaretnam Rajagopal, deputy director of Singapore's new bioinformatics institute.

existing bioinformatics centre at the National University and to help service the nation's expanding genomics programme. Its own research programme will follow from the interests of the staff who will be recruited internationally.

Gunaretnam Rajagopal, a theoretical physicist at Cambridge University, has been named as the institute's deputy director, and starts work in July. "I find the opportunity to build a world-class research organization, with strong encouragement, commitment and active support of the government of Singapore, irresistible," he says. ■

Rensselaer institute gets anonymous gift of \$360 million

San Diego In one of the largest donations ever made to a single university, the Rensselaer Polytechnic Institute at Troy, New York, has received an unrestricted gift of \$360 million.

The institute — one of the oldest technology colleges in the United States — will use the anonymous donation in a bid to double the scale of research activity and the size of its graduate programme, with particular emphasis on biotechnology and information technology.

The gift replaces an earlier offer of \$130 million made last December to pay for a Center for Biotechnology and Interdisciplinary Studies and an Electronic Media and Performing Arts Center. These facilities will go ahead as planned, but additional research, education and infrastructure projects will now follow.

Biotechnology research areas designated for support include functional tissue engineering, integrative systems biology, biocomputation and bioinformatics, and biocatalysis and metabolic engineering.

"This extraordinary gift will enable Rensselaer to play a leading role in enabling society to reap the still unimagined benefits

of these exciting technologies," says Shirley Ann Jackson, the institute's president.

Funding boost for women chemists in fight for equality

San Diego The Committee on the Advancement of Women Chemists (COACH), a group established to fight sexual discrimination in science, has received a \$1.4 million funding boost.

The US federal funds will further the group's attempts to break the 'glass ceiling' blocking women's careers. COACH will focus on identifying barriers, encouraging academic promotion and aiding recruitment.

The committee is chaired by chemist Geri Richmond of the University of Oregon, who formed the group with Jeanne Pemberton of the University of Arizona.

Ammunition radiation alone 'not a threat'

Brussels Radiation from depleted uranium ammunition "could not result in a detectable effect on human health", a panel of experts told the European Commission last week.

The report was commissioned after six cases of leukaemia were discovered among Italian soldiers who had served in Kosovo (see *Nature* **409**, 121; 2001). Uranium-238, a



Shell shock: radiation alone is not harmful, say experts, but added to toxicity it may be.

mildly radioactive by-product from nuclear power production, is used to strengthen the tips of shells. It vaporizes on impact, leading to fears that soldiers could have breathed in depleted uranium particles.

The expert panel reached its conclusions after reviewing available information and considering various exposure scenarios. The group noted that, although the radiation alone was not a threat, combined with depleted uranium's toxicity it might damage health — but there was no direct evidence to support suggestions that this had happened.

Canada invests heavily in research infrastructure

Montreal The Canada Foundation for Innovation, which supports research infrastructure projects, was last week

promised Can\$750 million (US\$480 million) of additional federal funding. This takes the total investment earmarked for the foundation to Can\$3.15 billion for the period to 2010.

The government-funded foundation was set up in 1997 to give infrastructure funding to universities and teaching hospitals. Grants given out so far total more than Can\$850 million. Matching funds from non-government sources have added Can\$1.2 billion on top of that.

The government also announced that four new "networks of excellence" will be set up. These will be devoted to research into automotive technology, genomics and society, clean water and early childhood development. Eighteen such networks already exist.

Bank to hand back impounded equipment

Barcelona Laboratory equipment impounded from the Immunology Institute of Bogotá (IIB) will be donated to a new research centre set up by the IIB's head Manuel Patarroyo, the management board of BBVA Banco Ganadero bank said recently.

The bank seized the equipment in January following non-payment of a debt by the hospital in which the institute was situated

(see *Nature* 409, 550; 2001). In mid-February, malaria vaccine researcher Patarroyo and his team decided to leave IIB and moved to a former Institute of Nuclear Affairs building. There Patarroyo set up the Immunology Foundation of Colombia.

The bank's president José Maria Ayala says the donation will take effect once Patarroyo confirms the constitution for his new centre.

Big NASA balloon manages just one day in space

Washington NASA officials were left scratching their heads again last weekend after the agency's second attempt to send a giant balloon to the edge of the Earth's atmosphere and around the globe ended in failure.

The Ultra Long Duration Balloon stayed aloft for 24 hours — an improvement on the few hours it managed last time (see *Nature* 410, 9; 2001). Pressure in the balloon had been expected to increase as the Sun rose and heated the gases inside it. But NASA scientists were forced to abort the mission when this did not happen. The balloon and its \$1 million payload eventually landed safely on the western coast of Australia, with just half a mile of dry land to spare.

Scientists are now analysing flight data to see if this replacement balloon had sprung a

leak, as the original did on its maiden voyage. The 60-metre-wide balloon is regarded as a potential alternative to satellites for data collection.

World's first webcam takes a coffee break

London Eight years old and 2.4 million visits in, what is reputed to be the world's first webcam is being retired. The site shows a coffee-pot in the University of Cambridge computer laboratories.

With only one coffee-pot serving several rooms, researchers became annoyed with trudging up flights of stairs only to find the pot empty. So they trained a camera on the pot and hooked the output up to the local network. Two years later, in 1993, it was linked to the fledgling World-Wide Web, where it built up an unlikely cult following. Now the site is being shut down as the labs move to new offices.

"It was only marginally more exciting than watching paint dry," admits Dan Gordon, one of the computer scientists behind the project. Catch it before it's gone.

► <http://www.cl.cam.ac.uk/coffee/coffee.html>

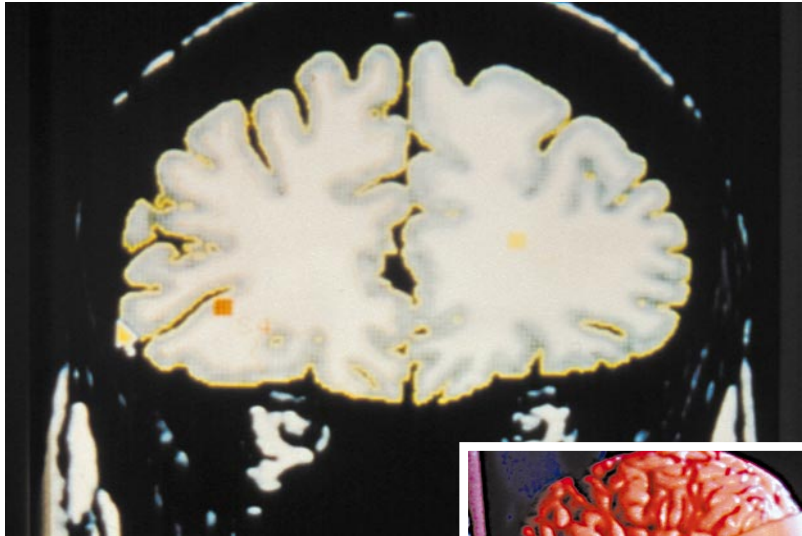


Get it while it's hot: website had many visitors.

Into the mind of a killer

Brain imaging studies are starting to venture into the legal minefield of research into criminal psychopathy. Alison Abbott reports from one of the most controversial frontiers of neuroscience.

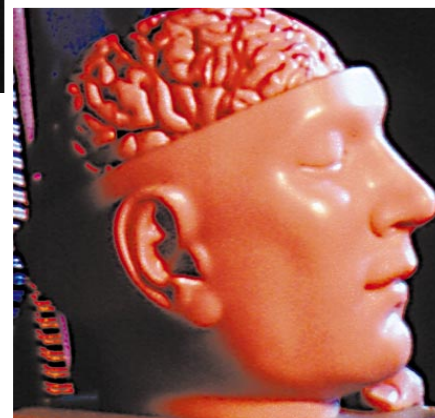
ADRIAN RAINE



Cesare Lombroso, the nineteenth-century Italian criminologist, was the first to argue on scientific grounds that criminals are born, not made. Drawing on emerging theories of evolution and genetics, and the contemporary fad for phrenology, he concluded that those with a 'criminal mind' could be identified by deformations of their skulls. It all seemed reasonable at the time. But the facts did not fit the theory, and Lombroso's research was later discredited.

Today, many psychiatrists accept that some people who fall foul of the criminal justice system suffer from a condition — psychopathy — that is as much an illness as, for example, schizophrenia. Environmental factors may help to determine whether this 'illness' is expressed in the form of violent, criminal behaviour, but a growing number of experts argue that the underlying condition is biological. "More and more data are leading to the conclusion that psychopathy has a biological basis, and has many features of a disease," says Sabine Herpertz, a psychiatrist at the RWTH-Aachen University in Germany.

Several researchers are now using the latest brain imaging techniques in an attempt to find out what is different about the brains of psychopaths. They hope that these studies will lead to a fundamental biological understanding of psychopathy, and perhaps even to drug treatments for the condition.



But other scientists fear that these researchers are unintentionally following Lombroso's path. They worry that it may not be possible to draw valid conclusions about a condition as complex as psychopathy from even the most carefully designed neuro-imaging studies. And they are conscious that the stakes for society may be high.

In no jurisdiction can a diagnosis of psychopathy currently be used to claim diminished responsibility. Quite the opposite: in the United States, prosecutors use such diagnoses as arguments against rescinding a death sentence. Throw brain imaging into the mix, and it might be seized upon as providing a simple marker of psychopathy — potentially biasing life-or-death judgments.

Psychopaths are not necessarily the sadistic killers of popular fiction. But they lack empathy, and are unable to experience guilt or remorse. They are assertive and egocentric, may be highly manipulative, and are



Life or death: will brain research save psychopaths from the chair, or send them to it?

unconcerned by the negative consequences of their actions. When they kill, the crimes are usually well planned and committed for personal gain. But engage in conversation with a psychopath, and he or she might seem perfectly normal.

Inhuman behaviour

Psychopaths form part of a wider group diagnosed as suffering from antisocial personality disorder (APD). They can be identified using a test developed by Robert Hare, a psychologist at the University of British Columbia in Vancouver^{1,2}. Diagnoses of APD are based primarily on patients' behavioural histories, but Hare's revised psychopath check-list (PCL-R) uses psychiatrists' assessments of the emotional deficits that underlie psychopathy.

An estimated three-quarters of prison populations fulfil the criteria for APD, and a quarter or more may exceed a PCL-R score of 30, the cut-off for true psychopathy^{3,4}. PCL-R scores also seem to predict the likelihood of reoffence⁵. Around four times as many psychopaths commit violence on release from jail as those simply diagnosed as having APD; and for psychopaths, the risk of reoffending is independent of the environment into which they return. Moreover, the higher the PCL-R score, the more likely they are to reoffend.

Being able to identify psychopaths using a check-list does not explain the condition's

FRANK SPOONER

biological basis. But long before the advent of brain-scanning techniques, psychologists were identifying physiological correlates of psychopathy that appeared to be linked to the underlying emotional deficit. Hare conducted many pioneering experiments in the 1960s. In some, for example, he exposed subjects to an unpleasantly loud noise after either a countdown or a visual stimulus. Over repeated trials, normal people became anxious as the countdown began or the visual stimulus was flashed up, and they began to sweat, increasing their skin conductivity. But psychopaths consistently showed low skin-conductivity responses⁶.

More recently, research has been extended to other physiological parameters. Every-one blinks when startled, and normally this response becomes more intense if they are placed in a threatening environment — which can be simulated in the lab by showing subjects a picture of an angry face. But in psychopaths, the startle response remains unchanged under these circumstances⁷.

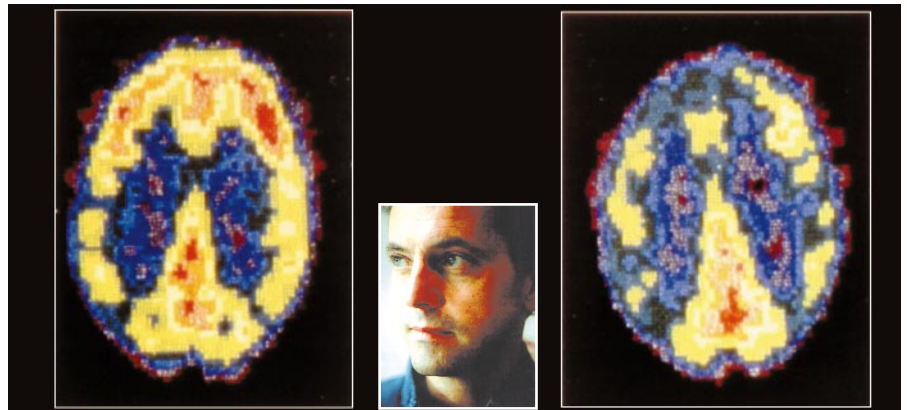
Psychopaths also respond differently to words that normally evoke an emotional response, such as 'kill', 'maim' or 'joy'. In tests in which people have to separate real words from nonsense words, normal volunteers process these emotionally charged words more slowly than neutral words, such as 'table' and 'butter'. But psychopaths process both categories of word at the same speed⁸.

Image consultants

The brain imaging techniques of positron emission tomography (PET) and magnetic resonance imaging (MRI) provide the opportunity to investigate psychopathy further. They might allow researchers to discover whether psychopaths' physiological and emotional deficits can be pinned down to specific differences in the anatomy or activation of the brain.

Among researchers who are starting to explore this area, there are two main theories of psychopathy. One, championed by Adrian Raine of the University of Southern California in Los Angeles and supported by the work of Antonio Damasio of the University of Iowa, gives a starring role to a brain region called the orbitofrontal cortex (see diagram, right). This is part of an area of the brain, known as the prefrontal cortex, involved in conscious decision-making. Damasio has shown, for example, that patients who have suffered damage in this area early in life have severe social behavioural problems and can be very aggressive⁹.

The other theory, promoted by James Blair of University College London, holds that the fundamental dysfunction lies within the amygdala, a small almond-shaped structure that plays a critical role in processing emotion and mediating fear. Recently, using PET scanning, Blair has shown that activation of the amygdala in normal volunteers is involved in responding to the sadness and



Mind games: Adrian Raine with PET scans of a normal brain (left) and that of a murderer. But, given the uncertainties about diagnosis, the significance of such work to psychopathy remains unclear.

anger of others¹⁰, and he hypothesizes that amygdala dysfunction could explain the lack of fear and empathy in psychopaths.

The two theories may not be mutually exclusive, Blair points out, as the orbitofrontal cortex, which does the 'thinking', and the amygdala, which does the 'feeling', are highly interconnected.

So far, evidence from neuroimaging studies is inconclusive. It is consistent with the theory that a deficit in emotional processing underlies psychopathy, as brain scans show a reduction in size, or changes in the activation of brain areas associated with emotional processing — including the prefrontal cortex, the amygdala, and other components of the 'limbic' system. But the findings do not promote one particular neurobiological theory.

One of the biggest problems is that many of the studies undertaken so far have used subjects who are not pure psychopaths. For example, Raine has used PET scans to compare the activity of the brains of murderers — both those who killed in a fit of rage, termed 'affective' murderers, and 'predatory' killers — with normal volunteers¹¹. He found anomalies in the scans of affective murderers, but not in those of predators. Some of the latter group may have been psychopaths, but none had been diagnosed using the PCL-R.

Last year, Raine's team published an MRI study of 21 men diagnosed as having APD, who admitted — under guarantee of confi-

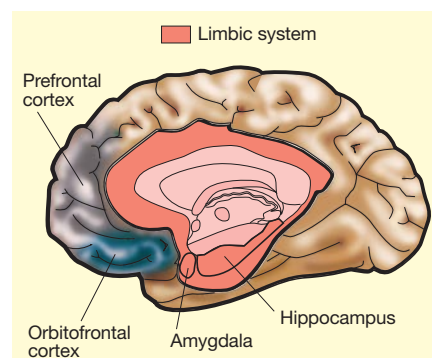
dentiality — to having committed violent crimes¹². Compared with normal volunteers, these men had shrunken prefrontal cortices — reduced in volume by up to 14%. The results grabbed the headlines, but their significance for true psychopathy remains unclear. Again, the volunteers had not been rated using the PCL-R, and shrunken prefrontal cortices are also associated with substance abuse — so the finding could merely indicate an association between drug-taking and APD.

Another study, published earlier this year by Mikko Laasko and his colleagues at Kuopio University Hospital in Finland¹³, used MRI to look at the volume of the hippocampus, part of the limbic system, in 18 male alcoholics who averaged a score of 31.2 on the PCL-R scale. There was a strong correlation between the PCL-R score and hippocampal shrinkage, but the conclusions were confounded because of the subjects' alcoholism.

Learned hostilities

Studies attempting to apply functional MRI (fMRI) to psychopathy are still in their infancy, and only one has been published so far. By monitoring blood flow within the brain, fMRI can study the activation of different brain areas at high resolution during the performance of specific tasks. Last year, psychologist Frank Schneider of the University of Düsseldorf and his colleagues described fMRI experiments in which 12 individuals with APD who scored highly on the PCL-R check-list, and 12 normal controls, were alternately exposed to a face with an emotionally neutral expression and a nasty smell¹⁴. This is a standard 'aversive conditioning' procedure, in which the expression on the face is perceived as more hostile as the experiment proceeds.

Both groups showed aversive conditioning, but in the APD patients activity was increased in both the amygdala and the prefrontal cortex while making the emotional association between the smell and the face. Superficially, this seems to argue against theories that a lack of activation in these brain areas is linked to psychopathy. But Schneider



Danger areas: psychopathy has been associated with dysfunction in several parts of the brain.

► suggests that the findings could indicate that psychopaths, because of their deficit in emotional processing, have to make a much greater effort to learn the associations forged in aversive conditioning experiments. But Schneider's sample included several subjects with a PCL-R score of less than 30.

Courting problems

The difficulty in interpreting such studies is adding to the unease of scientists who fear that the researchers involved are in danger of repeating Lombroso's mistakes. Claudio Luzzatti, a neuropsychologist at the University of Milan-Bicocca, argues that in other areas fMRI has muddled the waters, as its findings are often inconsistent with neuropsychological approaches. In studies of patients with brain damage, areas shown to be crucial for word comprehension and reading do not light up when these tasks are tracked by fMRI; whereas other brain regions, which had never been considered crucial, do. "It's not always easy to work out what fMRI is showing us," says Luzzatti.

John Marshall, a neuropsychologist at the University of Oxford, is similarly unconvinced by the studies conducted so far. "Criminality does not seem to be the kind of thing that can be localized," he says. "I am extremely sceptical about what people are actually measuring."

Hare agrees that neuroimaging studies of psychopaths should be interpreted with great caution. He adds that progress is hampered by the paucity of information about what the brains of various individuals look like when they are processing emotional information. "What we need are a set of standard tests for all sorts of clinical populations and normals," Hare says.

Scientists who have served as expert witnesses are concerned about how data from neuroimaging studies of psychopaths might be used in court. Helen Mayberg, chair of neuropsychiatry at the Rotman Institute in Toronto, says that PET scans are already being

used in an unscientific way by US defence lawyers — not to address the issue of psychopathy, but in a general attempt to convince courts to accept pleas of diminished responsibility. "A prisoner with no clinical diagnosis is taken for a PET scan — in shackles and maybe with a death sentence hanging over him," she says. "This is then compared with that of a non-shackled healthy person, and used in court as evidence of his state of mind at the scene of a crime committed years earlier."

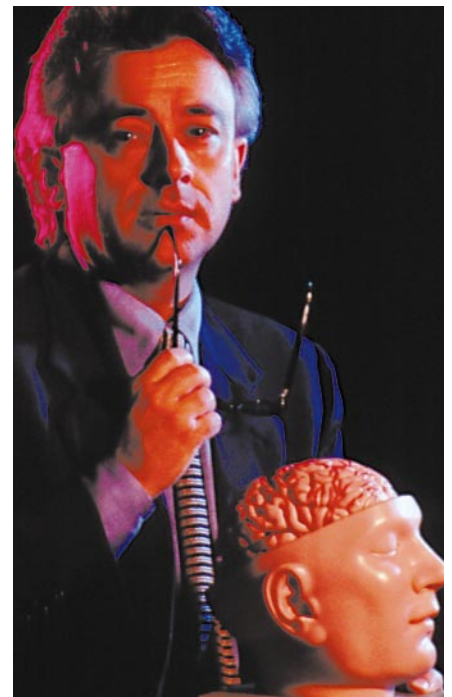
But proponents of neuroimaging research into psychopathy counter that such studies promise a more complete picture of psychopathy as a disease. In the long run, this could lead to changes in the law, allowing psychopaths to claim diminished responsibility and so routinely be sent to secure mental institutions rather than to prison. And as understanding of the neurobiological basis of psychopathy grows, new ideas for therapy may emerge. "The more we know about the biology of psychopathy, the clearer will be the path to the design of effective pharmacotherapy," argues Blair.

Dark side of the brain

Pharmacological approaches to treating psychotherapy would certainly be welcome. Behavioural therapy, in which psychopaths are put through programmes designed to help them change their behaviour patterns, has been spectacularly unsuccessful. For example, Michael Seto and Howard Barbaree of the University of Toronto followed the incidence of violent reoffending among 224 former prisoners who had undergone behavioural therapy before their release¹⁵. Those with high PCL-R scores, apparently doing well in therapy, actually reoffended more often than those scoring lower for psychopathy, who seemed to respond less well to behavioural therapy. Seto says that this result could reflect the manipulative nature of some psychopaths: the more psychopathic, the better they may be at convincing psychologists of progress under therapy.

One fundamental problem for psychopathy research is that there has historically been scant funding. "This is really the study of the dark side of human nature, and it finds few benefactors," says Hare.

But in Britain, that may be about to change. Following widespread concern that the criminal justice and mental health systems are failing to deal effectively with dangerous psychopaths, the government is planning fundamental reform. Most controversially, it intends to reform the Mental Health Act to allow "dangerous people with severe personality disorder" to be detained in secure mental institutions even if they have been accused of no crime. Although these particular provisions have alarmed civil liberties campaigners, the raft of measures also includes a major initiative within the prison service to improve the handling of those with APD — including psychopaths.



Thinking big: Antonio Damasio believes neuroscientists should investigate psychopathy.

Significantly, this initiative includes an annual £1 million (US\$1.5 million) research component, some of which may be used to support brain imaging and other neurobiological studies. One protocol under consideration, for instance, would follow the incidence of violent reoffense in released prisoners with high PCL-R scores and either high or low ratings on neurobiological indicators of emotional deficit.

Researchers who believe that neurobiological approaches offer the greatest potential for treating psychopathy hope that the British programme really will back the necessary research. "As it is clear that brain dysfunction can cause abnormal social behaviour, it is important for scientists to address the issue," says Damasio. "The human mind is complex and refined — but scientists should not be afraid to think big."

Alison Abbott is Nature's senior European correspondent.

1. Hare, R. D. *et al.* *Psychol. Assessment* **2**, 338–341 (1990).
2. Hare, R. D. *Manual for the Hare Psychopathy Checklist-Revised* (Multi Health Systems, Toronto, 1991).
3. Widiger, T. A. *et al.* *J. Abnorm. Psychol.* **105**, 3–16 (1996).
4. Hobson, J. & Shine, J. *Br. J. Criminol.* **38**, 405–515 (1998).
5. Hemphill, J. F., Hare, R. D. & Wong, S. *Legal Criminol. Psychol.* **3**, 141–172 (1998).
6. Hare, R. D. *Psychol. Rep.* **16**, 499–502 (1965).
7. Patrick, C. J., Bradley, M. M. & Lang, P. J. *J. Abnorm. Psychol.* **102**, 82–92 (1993).
8. Williamson, S. E., Harpur, T. J. & Hare, R. D. *Psychophysiology* **28**, 260–273 (1991).
9. Anderson, S. D., Bechara, A., Damasio, H., Tranel, D. & Damasio, A. R. *Nature Neurosci.* **2**, 1032–1037 (1999).
10. Blair, R. J. R., Morris, J. S., Frith, C. D., Perrett, D. I. & Dolan, R. J. *Brain* **122**, 883–893 (1999).
11. Raine, A. *et al.* *Behav. Sci. Law* **16**, 319–332 (1998).
12. Raine, A., Lencz, T., Bihle, S., LaCasse, L. & Colletti, P. *Arch. Gen. Psychiat.* **57**, 119–127 (2000).
13. Laasko, M. P. *et al.* *Behav. Brain Res.* **118**, 187–193 (2001).
14. Schneider, F. *et al.* *Neuropsychobiology* **42**, 192–201 (2000).
15. Seto, M. E. & Barbaree, H. E. *J. Interpers. Violence* **14**, 1235–1248 (1999).



No way back: Michael Seto has seen behavioural therapy consistently fail in psychopaths.

Fellowship fund would help eastern Europe to retain its young talent

Sir— Eastern European countries have a tradition of stressing the importance of solid scientific education and research. Even under communism, Polish scientists allowed to work in western laboratories were able to perform at the level of their colleagues from the host country.

This level of performance has been improving since the rise of democracy. Recently, leading western European laboratories — the European Molecular Biology Laboratory (EMBL) and the Max Planck Society — have created special PhD programmes to attract talented young scientists from eastern Europe. These programmes provide excellent opportunities for students to begin their scientific careers in well-funded western European laboratories. Unfortunately, they also have drawbacks. One is the risk of a flight of PhD candidates from eastern European laboratories such as those in Poland.

In Poland, there have been two major changes in the way laboratories are funded and operated since the transition to a democratic system of government. As a result, a new emphasis has been placed on the importance of graduate students in making a research programme successful. First, there has been the creation of a government grant scheme, awarding grants to researchers holding PhDs and *Habilitation* degrees — the qualification that allows a PhD to teach at university — on merit, as determined through a peer-review process. Second, there has been a movement away from the strict hierarchical structure which typified the traditional Polish laboratory. Instead of a team of researchers holding permanent positions and led by the head of the laboratory, contemporary Polish laboratories are associations of research groups consisting of principal investigators and graduate students working on the realization of research proposals.

Unfortunately, the government fellowships now offered in Poland to graduate students are inadequate to maintain a reasonable standard of living. As a result, principal investigators in Polish laboratories, facing strong competition from the West, will be unable to keep enough graduate students to perform research. Furthermore, there will be little incentive for the new generation of Polish scientists trained in western Europe to return to Poland and start their own laboratories, if they find a lack of qualified candidates for their research programmes.

What can be done to avoid this

pessimistic scenario? Until economic conditions in eastern Europe improve, one solution would be to allocate a small percentage of the funds used in current EMBL and Max Planck fellowship programmes to a fellowship fund that would allow eastern European students to work in their home countries. Western European institutions funding fellowships could monitor the PhD programmes by sending advisers to participate in committee meetings, examinations and public defences of PhD theses.

Not only would this help to preserve and promote good-quality research and scientific training in eastern European countries, but it would also create a platform for collaboration between researchers in western and eastern Europe.

Jarosław Marszałek, Krzysztof Liberek, Igor Konieczny

Department of Molecular and Cellular Biology, Faculty of Biotechnology, University of Gdansk, 24 Kladki, Gdansk 80822PL, Poland

Arabidopsis could shed light on human genome

Sir— Although only recently published in *Nature*¹, the entire sequence of the model plant *Arabidopsis thaliana* has been in the public domain for a considerable length of time. We note, however, that the papers analysing the data of the human genome substantially lack comparisons to the *Arabidopsis* genome.

Unlike any of the animal genomes sequenced to date, the *Arabidopsis* genome is virtually without significant gaps, and even includes a substantial amount of heterochromatic sequence which has been of great use in examining centromere structure and function².

To cite but a single comparison arising from our own work, we have identified the entire complement of SNAREs in *Arabidopsis*³. Most notably, this mere slip of a plant encodes for more SNAREs (53, ref. 3) than humans do (35, ref. 4), including some classes of SNAREs which are not found in either human or yeast cells³.

We can understand that there is limited space for such analyses, but surely one should not omit significant information pertaining to an entire kingdom of eukaryotes!

If a comparison with yeast genes is of relevance to the evolution and function of the human orthologues, would not a comparison with a more distantly related and multicellular eukaryote reveal even more?

Anton A. Sanderfoot, Natasha V. Raikhel

Department of Energy Plant Research Laboratory,

Michigan State University, East Lansing, Michigan 48824, USA

1. The Arabidopsis Genome Initiative *Nature* **408**, 796–815 (2000).
2. Copenhaver, G. P. *et al.* *Cell* **100**, 367–376 (2000).
3. Sanderfoot, A. A. *et al.* *Plant Physiol.* **124**, 1558–1569 (2000).
4. Bock, J. B., Matern, H. T., Peden, A. A. & Scheller, R. H. *Nature* **409**, 839–841 (2001).

Climate panel looked at all the evidence

Sir— I read with interest your News article (*Nature* **409**, 445; 2001) covering the approval of the Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) Working Group I in Shanghai in January. As a coordinating lead author of one of the chapters, lead author of the technical summary and a member of the drafting team of the summary for policy-makers, I would like to clarify two points.

First, although climate modelling has advanced during the past five years, this is not the main reason for the revised range of temperature projections. The higher estimates of maximum warming by the year 2100 stem from a more realistic view of sulphate aerosol emissions. The new scenarios assume emissions will be reduced substantially in the coming decades, as this becomes technically and economically feasible, to avoid acid rain. Sulphate emissions have a cooling effect, so reducing them leads to higher estimates of warming.

Second, it is incorrect to say that the report does not consider new research on a possible change, or even shut-off, of the North Atlantic thermohaline circulation. This important new insight is addressed and assessed in detail in chapters seven and nine, amplified in the technical summary of Working Group I and explicitly carried over to the summary for policy-makers. All relevant and recent publications are cited.

Thomas Stocker

Physics Institute, University of Bern, Sidlerstrasse 5, 3012 Bern, Switzerland

Right person, wrong sex

Sir— Although the enthusiasm of research institutions to showcase the important contributions from women scientists is understandable, I hope it hasn't extended so far as to change the gender of Clair Patterson, the man who gave us the age of the Earth — but was referred to in a book review (*Nature* **409**, 20; 2001) as “she”.

William H. Casey

Department of Land, Air and Water Resources, University of California–Davis, 1 Shields Avenue, Davis, California 95616, USA

Nature replies — Apologies: the mistake was ours.

Opening a window on space

Advances and discoveries that will help us to find our place in the Universe.

New Cosmic Horizons: Space Astronomy from the V2 to the Hubble Space Telescope

By David Leverington

Cambridge University Press: 2000. 520 pp.
£55, \$85 (hbk), £19.95, \$29.95 (pbk)

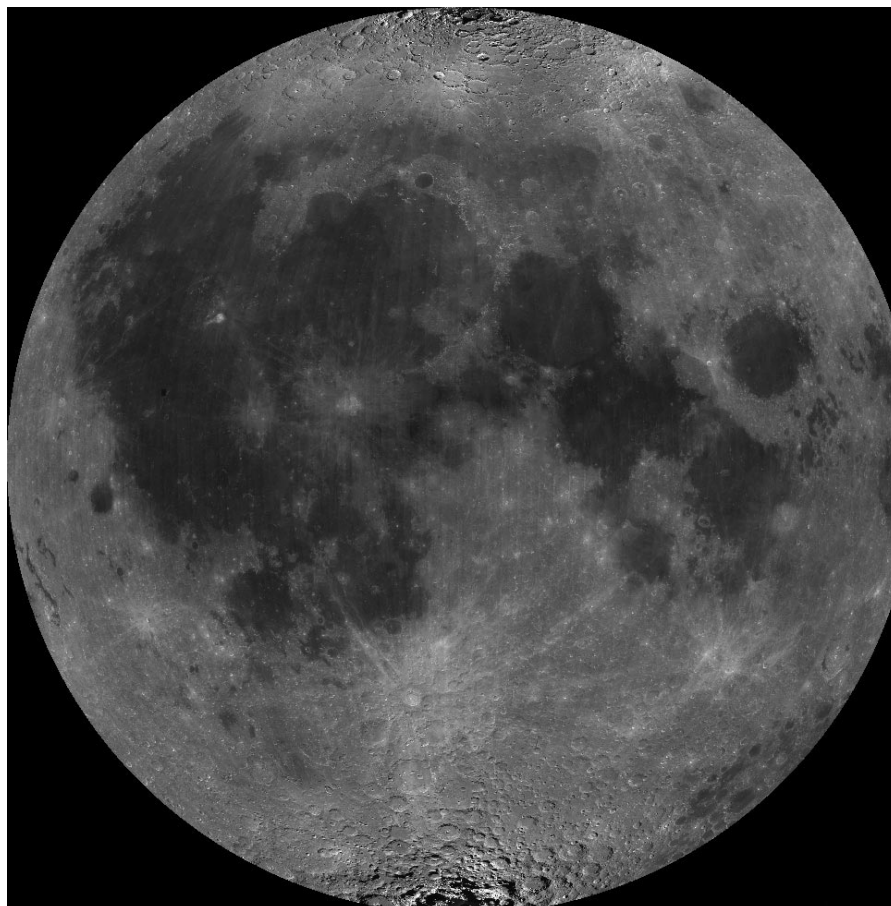
G. Siegfried Kutter

To appreciate the story of space astronomy, we must go back to the years just after the Second World War. Astronomers then were largely limited to observing the night sky in optical light, using ground-based telescopes — although a few visionary types had been searching for cosmic radio waves since the early 1930s. By the mid-1940s, for example, astronomers knew that the Sun did not occupy a preferred position, but was thought to lie perhaps as much as 30,000 light years from the centre of the Milky Way. And the Milky Way had not yet been convincingly identified as a spiral galaxy.

It was known that galaxies were distributed roughly uniformly in all directions, when averaged over sufficiently large scales, and, most interestingly, that all were receding from one another. These discoveries led to two competing cosmological theories: the big bang theory, proposed in the 1920s, and the steady-state theory, published in 1948. The former claimed that the Universe came into existence a finite time ago in a gigantic explosion causing the observed cosmic expansion. The latter claimed that on sufficiently large scales the Universe is never-changing and that as galaxies recede from one another new matter continuously comes into existence, coalesces into stars and galaxies, and fills the voids created by the cosmic expansion.

Few astronomers realized then the renaissance their field was about to experience. This had its origin in the technologies developed during the war and in the East–West competition the war precipitated. Rocketry, radar and computers allowed engineers and physicists to send telescopes and detectors above the terrestrial atmosphere, which blocks out most electromagnetic radiation — γ -rays, X-rays and most ultraviolet and infrared radiation. The cold war ensured that ample resources were available to develop these technologies further and to go into space.

The story of how the terrestrial window was opened to the full range of cosmic electromagnetic radiation, and the impact this had on the development of astronomy is the subject of David Leverington's book. He starts with the first flights of V2 rockets from the New Mexico desert and ends some 50 years later with NASA's Hubble Space Telescope. Along the way he covers the Soviet



Nearest neighbour: but the race to the Moon was littered with failed missions.

Sputnik and American Explorer projects, the race to the Moon, missions to the planets, space-based solar research and space observatories dedicated to research beyond the Solar System. He tells of competition and cooperation, of successes and failures. He describes the establishment of national and international institutions and the hard financial, political and technical decisions that were made. And he offers background material on the history and purpose of each mission to place it in the broader context of national and international space agendas.

The book's layout is chronological. After describing individual missions, Leverington focuses on their major scientific results, giving frequent cross-references to other missions and, as appropriate, the contributions made by ground-based and airborne astronomy. The result is a reasonably complete overview of the contributions space research has made to astronomy.

For instance, we learn that the first major and totally serendipitous scientific discovery of the space age was by James Van Allen in 1958, when Geiger counters he had placed on

two Explorer spacecraft detected radiation from charged particles trapped in the Earth's magnetic field. Van Allen had discovered the inner of two radiation belts around the Earth, now named after him.

In describing the race to the Moon, Leverington points out that sending spacecraft to the Moon was much more challenging than launching Earth-orbiting satellites. Many of the early lunar missions failed because the spacecraft exploded, suffered ignition failure, fell back to Earth or followed flawed trajectories that missed the Moon by tens of thousands of kilometres. And Leverington describes the first large international mission of the space age — the coordinated effort by Western Europe, Japan, the Soviet Union and the United States to intercept Comet Halley during its 1985–86 visit to the inner Solar System.

After covering planetary and solar missions, Leverington turns to observations of sources outside the Solar System. In his discussion of the high-energy sky (that is, cosmic γ -ray, X-ray and UV sources), he starts with the discovery in 1962 of the first extra-

book reviews

solar X-ray sources, and continues with the increasingly sophisticated missions of the 1970s and early 1980s (for example, the Exosat mission) and those of the late 1980s and 1990s (such as Rosat and the Extreme Ultraviolet Explorer).

The development of detectors for infrared radiation lagged behind those for X-rays and γ -rays. Thus, the first infrared astronomical spacecraft, the Dutch–UK–US Infrared Astronomy Satellite (IRAS), was not launched until 1983. It was followed during the next dozen years by NASA's Cosmic Background Explorer (COBE), Hubble with its infrared capabilities and the European Infrared Space Observatory. Perhaps the most significant result of infrared space astronomy to date has been COBE's precise determination of the temperature of the cosmic microwave background radiation. This radiation is thought to be the cool afterglow of the Universe's explosive beginning, and lends powerful support to big-bang cosmology. It makes the steady-state theory virtually untenable.

In his final chapter, Leverington discusses the Hubble Space Telescope from its roots in a classified 1946 RAND study. This study considered possible scientific programmes using optical space telescopes of up to 15 metres in diameter — rather ambitious thinking for the time. But it was not until the early 1970s that NASA began to design Hubble seriously, and the telescope was not launched until 1990. Leverington divides his coverage of Hubble's major results into the periods before and after the first servicing mission in 1993, which corrected the telescope's flawed primary mirror. He describes many of the results that have made headlines over the years, such as the discovery of gravitational lensing by foreground galaxies of more distant galaxies; the Hubble Deep Field images showing a bewildering variety of very faint galaxies in the early stages of formation; and a determination of the value for the Hubble constant, H_0 , which corresponds to an age for the Universe of between 9 and 15 billion years depending on the values chosen for various cosmological parameters.

Leverington evidently did his homework. Not only is the writing in *New Cosmos Horizons* thorough and clear, but the book is also well structured, and the appendices, glossary, bibliography and other summaries allow the reader to navigate it reasonably easily. Besides, as a participant in space research since the 1960s, the author witnessed many of the events first-hand, giving his writing authority.

There are a few omissions. Except for two fleeting references to the Space Telescope Science Institute, Leverington discusses none of the science centres for the missions he describes, even though these centres are as much a part of the modern space observatories as are their funding authorization, development, construction and launch. And the book's coverage ends in 1996. Leverington explains that this allowed him to gain some

historical perspective on the missions and results he presents. A summary of more recent missions, however, and of projects now in the design stage or under development, would have broadened the historical context of the book.

But these are minor shortcomings. I recommend the book to anyone interested in space astronomy — laypersons, students of astronomy, and professional astronomers. Besides learning a good deal about technology and science, they will come away with the appreciation that progress in space astronomy is a broadly social effort, involving, besides universities and other research institutes, individuals of vision and courage, armies of engineers and technicians, industrial companies, national and international agencies, the ebb and flow of competition among nations and, ultimately, the good will of governments and taxpayers to provide its funding. ■

G. Siegfried Kutter is at TRW, Inc., PO Box 6759, Breckenridge, Colorado 80424, USA.

Journey of a discovery

Summon up the Blood: In Dogged Pursuit of the Blood Cell Regulators by Donald Metcalf

AlphaMed Press: 2000. 214 pp. \$49.95 (hbk), \$29.95 (pbk)

Fred S. Rosen

Outside the door of my office, Don Metcalf's name is inscribed on a plaque festooned with the names of more than a dozen other men (alas, no women!) who have received the Warren Alpert Foundation Prize for a "signif-

icant contribution to the prevention, treatment or cure of a major disease". In his book, Metcalf recounts the tale of the discovery for which he received the prize, that of colony-stimulating factors (CSFs) — protein molecules that regulate the formation of white blood cells in the bone marrow. Metcalf's narrative reads like an account of the Antarctic explorer Ernest Shackleton's 800-mile boat journey in search of help after his ship was crushed by ice — the achievement of a goal on a distant shore, reached by dead reckoning in a storm-tossed sea against overwhelming odds, a tale of incredible human endurance.

With his deep knowledge of the biology of bone-marrow progenitor cells, but using cumbersome, imprecise bioassays, Metcalf and his numerous co-workers managed to purify CSF. Prompted by the discovery of erythropoietin — the hormone that stimulates the growth of red blood cells in the bone marrow — they set out with the belief that an analogous growth factor must exist for white blood cells. They thought that this putative substance might have a role in the pathogenesis of leukaemia.

Metcalf and his team fractionated enormous volumes of urine and tissue to obtain nanograms of active material, too little to determine the protein's amino-acid sequence. Their assays involved growing progenitor cells in soft agar, and it took days to get a result. Many batches of plastic culture dishes and fetal calf serum had to be discarded because they were found to be toxic to the cultures. And there was a continuous search for a cell source to provide a 'feeder layer' for the tissue cultures.

Finally, Metcalf's profound insight into the biology of progenitor cells led to the discovery that there were four types of CSF — now called GM-CSF, M-CSF, G-CSF and



A positive culture: the road to success developed the careers of many scientists in Metcalf's lab.

REPRINTED WITH PERMISSION FROM ALPHAMED PRESS

multi-CSF, where G and M represent respectively the white blood cells known as granulocytes and monocytes.

The road to success also developed the careers of many outstanding scientists in Metcalf's laboratory. This book is a tribute to the 329 researchers who came to the Walter and Eliza Hall Institute of Medical Research in Melbourne to work with him over the course of more than three decades. Scattered throughout the book are photographs of these young scientists, posing with Metcalf — whose hair becomes greyer, his body stouter and his demeanour gaining in distinction during the course of those years.

The title of the book is a quote from Shakespeare's *Henry V*, spoken on the eve of the Battle of Agincourt. But the "dogged pursuit of the blood cell regulators" turns out more to resemble the war of attrition that was the prolonged Battle of Passchendaele rather than the victory of Agincourt. Or better yet, the lament of the convicts who reached Australia's "Fatal Shore". The countless hours spent peering down a microscope gave Metcalf severe back pain; we learn that he eventually had to have a spinal fusion. And there were the scientific competitors who, like relatives, are noxious people thrust upon you rather than those with whom you might choose to interact. Then came the greedy American biotechnology companies, which wagered and won large sums of money in the industrial development of the CSFs before there was any solid evidence that they might have therapeutic benefits. Finally, it turned out that CSFs probably have little or nothing to do with the pathogenesis of leukaemia.

Yet still there was victory, because the fact remains that the CSFs have been of great therapeutic value, particularly G-CSF, for elevating white blood cell counts. The difficult labour had a happy outcome, and we can forgive Metcalf his sardonic view of the world and of his fellow scientists at the Walter and Eliza Hall Institute. As Metcalf points out, the full therapeutic benefits of CSFs have yet to be exploited; it takes decades to realize the value of such discoveries. Now, more than 50 years after the serum protein γ -globulin became commercially available, new uses for it are being found. And 25 years after their discovery, monoclonal antibodies are just beginning to fulfil their therapeutic potential.

Certainly, Metcalf's discoveries deserve a Nobel prize. But that reminds me of a story John Enders told me about his 1954 Nobel prize, awarded for cultivating the polio-myelitis virus. He immediately called his father, a banker, to tell him of the news, and his father replied: "Whoever heard of getting a prize for doing your job?" ■

Fred S. Rosen is at the Center for Blood Research, Warren Alpert Building, 200 Longwood Avenue, Boston, Massachusetts 02115, USA.

A microcosm of society

Utopia's Garden: French Natural History from Old Regime to Revolution

by E. C. Spary

University of Chicago Press: 2000. 338 pp.

\$70, £44.50 (hbk), \$25, £16 (pbk)

Matthew Ramsey

The botanical garden established in Paris in 1626 by King Louis XIII of France was originally intended for the cultivation of medicinal plants. But this was to be among the least of its many functions. To the city's population the Jardin du Roi offered a magnificent park, the first public garden in the French capital. Successive governments found other uses for it. During the French Revolution, for example, the garden's nurseries produced patriotic 'trees of liberty' that

were planted in front of public buildings.

The world of science, for its part, knows the Jardin des Plantes as the site of the Muséum National d'Histoire Naturelle. When the revolutionary National Convention founded the Muséum in 1793, it ratified and reinforced the garden's already prominent role as a teaching and research centre with impressive natural history collections. The garden owed its scientific distinction chiefly to Georges-Louis Leclerc, comte de Buffon, the great naturalist who served as its superintendent from 1739 until his death in 1788. By naming the new institution the Muséum National, the Convention conferred a suitably republican title on the former Jardin du Roi and emphasized that the garden itself was but one part of this larger undertaking.

Emma Spary appropriately makes the garden the focus of her study. This strategy both recognizes its pivotal role in the development of French natural history and suits Spary's approach to the history of science. Drawing eclectically on a variety of influences—Steven



Burgeoning bower: Louis XIII's garden survived the French Revolution to flourish under the Republic.

Shapin, Simon Schaffer, Bruno Latour and Michel Foucault, among others — her account treats natural history as a set of inter-related practices that must be understood within their larger historical context. Her complex argument intertwines three major lines of analysis: nature, politics and society.

First, Spary describes the garden's place within the patronage system of the Old Regime (pre-revolutionary France). The superintendent owed his authority and resources to the favour of crown and court; his subordinates, in turn, were in a sense his personal dependants. Buffon was both a recipient and a dispenser of patronage, and his success as an institution-builder hinged on his mastery of these complementary roles. Spary then recounts the garden's transformation under the revolutionary Republic, at a time when claims to authority required a new, democratic foundation. The garden's survival and rebirth are a rare success story — the revolutionaries condemned the old scientific and learned institutions as bastions of privilege, and abolished nearly all of them.

Secondly, Spary emphasizes the place of humans in eighteenth-century natural history. Man was at the very centre of the natural world, not set apart from it. The boundary between nature and society became blurred, as did the distinction between the natural and social sciences. Spary is particularly good at describing the ramifications of their shared vocabulary: constitution, economy, culture, degeneration, regeneration, naturalization, *mœurs* (a hard-to-translate French term meaning morals, manners, customs). Nature could serve to justify or criticize social and political institutions; conversely, our understanding of society could illuminate natural phenomena. A key example is Buffon's conception of liberty. In nature, a species cannot flourish unless it is free to choose favourable conditions, including a suitable climate with adequate food sources. This principle had its social parallel in the idea of liberty as a fundamental human right. But nature imposes limits. Not all climates and conditions are hospitable to life, and so, Buffon argued, species tend to degenerate rather than improve.

Here mankind does have a unique role to play. Our skilled intervention can benefit both non-human species and our own — the first through horticulture and animal husbandry, the second through economic and social management. This was Rousseau turned upside-down. Whereas, for Rousseau, civilization warped human nature, the naturalists maintained that civilization, enlightenment and scientific expertise would serve to perfect all species, including our own. "Utopia's Garden" is neither an Eden from which we have been expelled, nor a primal state of nature that we have left behind. Rather, it is a work of art and science that we must construct; the Jardin des Plantes was the microcosm of a much larger practical enter-

prise. This confident promise had a powerful appeal for revolutionaries bent on regenerating the French nation.

There is much more to be gleaned from this erudite, wide-ranging and perceptive study than any brief account can possibly convey. It may take some patience to harvest its riches, particularly for readers unfamiliar with the contributions of science or the events of the French Revolution, which are not systematically explained. Specialists may carp at the sometimes less than rigorous use of theoretical constructs, certain simplifications (of revolutionary ideology, for example), or the occasional historical inaccuracy, as in a muddled paragraph on the collapse of the monarchy in August 1792. But these are minor concerns. Those who persevere will be handsomely rewarded. ■

Matthew Ramsey is in the Department of History, Vanderbilt University, Nashville, Tennessee 37235, USA.

When it came to naming names

Neurological Eponyms

edited by Peter J. Koehler, George W. Bruyn & John M. S. Pearce
Oxford University Press: 2000. 400 pp.
£39.50, \$54.95

Roy Porter

Those familiar with Barry Firkin and Judith Whitworth's huge *Dictionary of Medical Eponyms* (Parthenon, 1989) will be aware that some branches of medical science gather more eponyms than others, and neurology is one that gathers a lot of them. This is for the same reason that so many capes and creeks in Africa are named after Victorian explorers. Around the time that eponym-conferral was the done thing, discoveries aplenty were still there for the making in neuroanatomy and neuropathology, as they were in the geography of the Tropics. Cynics might add that neurology in those days was such a recondite, marginal and arduous field of inquiry that prodigality in naming helped the neurological community to keep its spirits up.

Certainly, one suspects that, insofar as figures such as Froment, Moro, Adie and Chiari are remembered at all these days, it is by the good fortune of having had (respectively) a sign, reflex, syndrome and malformation named in their honour, if honour it be. And this well-researched and well-organized volume, written by an international team of experts, will tell you all you need to know about those six figures and about some 50 others commemorated in the neurological pantheon through a named disease, structure, process or symptom. (William John Adie, in case you are as ignorant as I was until recently,

was a "Geelong boy", who did heroic service in the Great War, carried out research in London, and is immortalized through his investigations into abnormal kinds of pupil dilation and the underlying paralysis of the iris sphincter that causes them.)

Alongside the relatively unsung, this book quite properly celebrates the top neurologists and their signs and syndromes. This begins, appropriately enough, with Adamkiewicz's artery. There follow the circle of Willis, Head's areas, Alzheimer's disease, Down's syndrome, Huntingdon's chorea, Parkinson's disease, and many more.

Embedded in accomplished analyses of their education, research careers and scientific reputation, less familiar facts are often to be found. The eponymous James Parkinson was, of course, a London general practitioner active during the reign of King George III, and a pioneer of the infant science of palaeontology. Not all will know, however, that he was also a political activist and pamphleteer. And he was a member of the London Corresponding Society, a radical club dedicated to furthering the rights of man. Parkinson is perhaps the only contributor to neuropathology to have been arrested for alleged involvement in a plot to murder a reigning monarch.

What is disappointing in this book is that so few of the essayists chose to explain precisely how proper names got attached to the various parts, processes and pathologies. Frank Clifford Rose tells us, for example, that it was the French neurologist Jean-Martin Charcot who called Parkinson's 'shaking palsy' *la maladie de Parkinson* — although it may be added that Charcot then went on to stake his own claim to be the definitive codifier of the disorder.

But in most cases we are left in the dark. Was it mainly disciples who affixed the label? What was in it, we might ask, for the nominator? And did not endless priority and property disputes then flare up, which would have made fascinatingly unedifying reading?

We are told in passing, for instance, that the "notoriously chauvinistic French" (another eponym, incidentally, from Nicolas Chauvin, an old Napoleonic soldier) talked about the Claude Bernard syndrome, instead of Horner's. We also find out that a French professor of neurology, Jean Barré, coined the term 'Adie's syndrome' for the same syndrome, but then, perfidiously "running true to French style", changed his mind and proposed 'Weill-Reys syndrome'. Worse still, the Germans used the "blatantly false eponym" Kehrer-Adie syndrome. Poor old Horner!

But all this comes in scraps. What a shame that the machiavellian workings of linguistic imperialism in medical science are not addressed head-on in a book that, for all its merits, fails to see the historical wood for the trees. ■

Roy Porter is at the Wellcome Institute for the History of Medicine, 183 Euston Road, London NW1 2BE, UK.

Questions of direction

Top-down versus bottom-up.

David Lindley

A few years ago, I had a conversation with an editor about the possibility of writing a book on global warming and climate change. There had been a lot of stories in the newspapers about new measurements of global average temperature, the difficulty of establishing trends, the lack of a good understanding of climatic and atmospheric changes in the Earth's past, and so on. No one really knew what was going on and what the implications were. I could write a book laying out the scientific arguments with all their uncertainties and unknowns, and showing why this was such a confusing subject.

But confusion is not a big selling point. The editor wanted to know whether the planet was really getting warmer or not. Probably, I said, but it's not clear how quickly. This was not an attractive answer. I got the impression that this editor would be interested in a book that said either (a) global warming is a huge hoax, there's nothing to worry about, or else (b) New York City will be under a metre of water in 30 years' time. Conspiracies and disasters have market potential.

I never wrote the book. It probably would have been a little dry. Useful and illuminating, no doubt, but how many people really want to read a dispassionate and thorough account demonstrating that a complex issue is, in fact, truly complex? That's what university presses are for.

It's a trite complaint that journalists and science writers oversimplify, turning all shades of grey into black or white. It makes for a clearer message and (perhaps) it sells better. But scientists themselves are just as apt to promote seemingly simple answers to complicated questions, in the interests of getting their point of view across.

At the heart of Stephen Hawking's *Brief History of Time*, for example, is the idea that fundamental physical laws can determine the whole course of cosmic history, with the particular proposal that something called

Evolution may be at heart a simple idea, but its ramifications are almost infinitely complex.

the 'no boundary condition' can sweep a lot of uncertainties under the carpet. Richard Dawkins has pushed the idea of "selfish" genes, so that even apparently altruistic acts are to be explained as the result of one collection of genes improving its chances by assisting in the survival of other bodies carrying the same genes. I remember how appealingly powerful Dawkins made this idea sound, yet I couldn't help thinking that the wool was being pulled over my eyes. How, after all, does one hyena know that some other hyena is its second cousin, so it can calculate the degree of risk it's willing to take protecting the other animal?

Simple ideas with apparently far-reaching implications seem to hold out the hope of resolving difficult and murky questions. Why is the Universe the way it is? Why do animals do the things they do?

On the other hand, as Oscar Wilde warned, "the truth is rarely pure and never simple". Science may, on occasion (and perhaps in theoretical physics above all), generate simple, elegant explanations of vast power and reach. But this is not necessarily the rule, even in physics, nor a good model for science in general. Evolution, for instance, may be at heart a simple idea, but its ramifications are almost infinitely complex, and it certainly doesn't lead to simple predictions in the way that general relativity does.

Whether one prefers science to be simple or complicated may be essentially a matter of psychology. Edward O. Wilson, in his recent book *Consilience: The Unity of Knowledge*, argues that the disparate areas of science are fundamentally of a piece, and moreover that science can in principle offer rational accounts of all manner of seemingly unscientific things, extending even to artistic creativity. Wilson is eloquent and passionate as always, but the world he conjures, although compelling to many, seems dreary to me — the intellectual equivalent of the globalization of human culture.

There may be an unbridgeable divide between those who hanker after grand systems and those who revel in detail, complexity and even ambiguity. As a physicist turned writer, I seem to be moving from the first camp to the second. I don't think imposition of an arbitrary 'no boundary condition' can explain all that much about the Universe. I think there's more to life than the selfishness of genes. I don't think Wilson will ever be able to furnish a scientific explanation of why, the older I get, the more I like Haydn.

A recent trend in popular science publishing suits my changing philosophy.



Dava Sobel's *Longitude* is the archetype. The ability to sail around the world without crashing into rocks is a big theme, of course, but the attractions of the book are in the story-telling and the accumulation of incident. A couple of recent books on the mathematical concept of 'zero' work in somewhat the same way, tackling a big subject by means of the sharp detail and the telling anecdote.

This, it seems to me, is not only a more interesting way of writing about science, but a more honest and effective one. Science grows in a piecemeal, haphazard fashion more often than it does by blinding flashes of insight, and writing that reflects the erratic and zigzag growth of scientific knowledge is ultimately more persuasive and enlightening.

So perhaps that book on global warming could be written after all. Delve into the history of weather records, from the obsessive diaries of eccentric country curates to today's world-spanning technology. Show, don't tell, how the big picture arises by the aggregation and resolution of often contradictory minutiae. Make plain, above all, why scientists can be both knowledgeable and uncertain. ■

David Lindley is a writer currently living in Arlington, Virginia, USA.

CORBIS

Genesis by definition

Robert W. Cahn

A small subdivision of that flourishing discipline, the history of science, is the pursuit of the genesis of scientific disciplines. The central difficulty of that pursuit is to define a discipline, or at least delineate an envelope that can contain the essentials of a discipline.

Whenever I contemplate that recalcitrant requirement, I am reminded of Aaron Katchalsky, who said of biophysics (a modern discipline he had helped to create) that “biophysics is like my wife. I know her, but I cannot define her”.

To exemplify the problem of definition, consider what the *Oxford English Dictionary* (OED, 1989 edition) says about physics: “The science, or group of sciences, treating of the properties of matter and energy, or of the action of different forms of energy on matter in general (excluding chemistry, which deals specifically with the different forms of matter, and biology, which deals with vital energy).” ‘Physics’ derives linguistically from *ta fusika*, meaning ‘natural things’, so it might equally well be a name for biology (which certainly deals with more than vital energy).

In attempting to specify what my own modern discipline, materials science, is about, I increasingly had recourse to what philosophers call ostensive definition — to point at something and say: “That is a specimen of materials science.” One intention for the *Journal of Materials Science*, founded in 1965, was that its contents should, in the long

run, constitute an ostensive definition of this then-new subject. Those who submitted papers to the journal presumably thought — and think — of themselves as materials scientists. I believe that the journal’s contents 35 or so years later demonstrate unequivocally that the progressive ostensive definition has worked.

When one consults the OED for a definition of ‘discipline’, one finds: “A branch of instruction or education; a department of learning or knowledge; a science or art in its educational aspect.” The linchpin is that adjective, educational, for indeed, the identification and survival of disciplines is nowadays determined in universities.

Historically, individual disciplines, for example physical chemistry, chemical physics (hard to distinguish these two!) and colloid science, as well as my own, have had to withstand successive stages of ridicule and takeover attempts before becoming established or accepted. Perhaps it is through the harsh trial of academic infighting that disciplines win their spurs. Materials science has made it, colloid science never quite did.

The modern ‘social’ approach to the history of science produces its own approach to an understanding of ‘disciplines’ in general. John Ziman, a Bristol University physicist turned commentator on science in general, writes: “An academic discipline is much more than a conglomerate of university departments, learned societies and scientific journals. It is an ‘invisible college’ whose members share a particular research tradition ... A recognized discipline or sub-

Discipline

“The ‘social’ approach to the history of science has need of prompters, authority figures who whisper advice so active protagonists don’t stray too far from the bounds of the discipline.”

discipline provides an academic scientist with a home base, a tribal identity, a social stage on which to perform as a researcher.” One might go on to say that this social stage has need of prompters, established authority figures who whisper advice from the wings so that active protagonists do not stray too far from the currently accepted bounds of the discipline in question. But then, the best actors do not need prompters.

To my mind, there are two large families of disciplines which differ in the form of their genesis. Physical chemistry, emerging around 1880 from organic chemistry, was a case of emergence by splitting. In this type of discipline birth, the splitting does not always ‘take’: thus solid-state physics, held in vociferous contempt by Wolfgang Pauli when it was quite new 70 years ago, is now pursued by many who would not consider it a discipline distinct from ‘physics’.

The gradual emergence of geology to amalgamate petrology, stratigraphy, orography, mineralogy and palaeontology, among others, is a case of emergence by integration. So a modern geologist, like a materials scientist, is apt to be quite thinly spread if he or she takes the breadth of his or her science seriously.

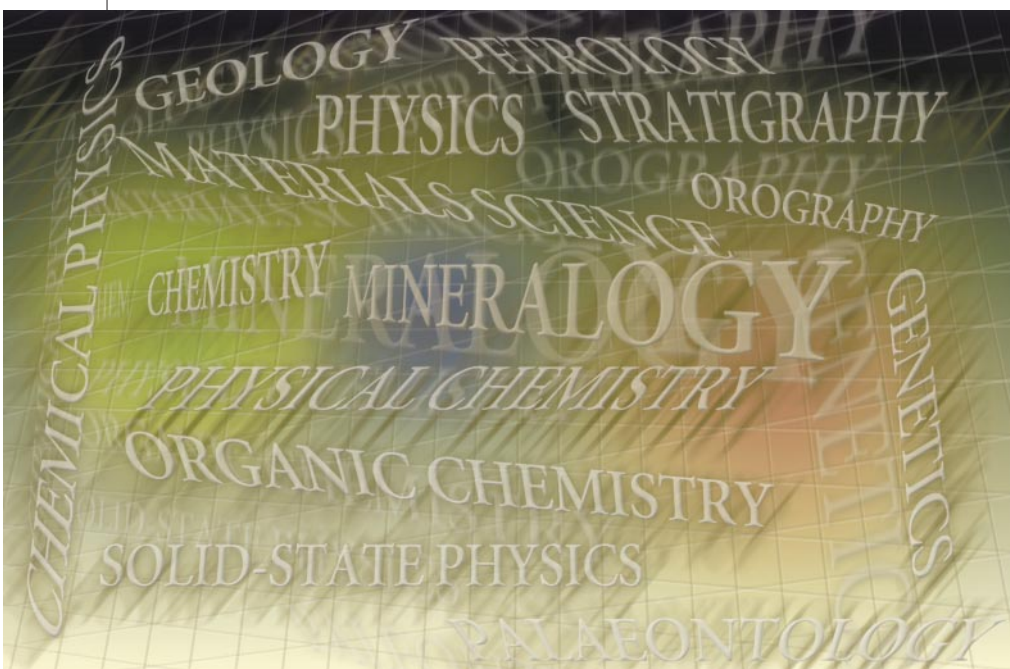
A view held by some is that a discipline is only ‘real’ if a single founding parent can be identified, such as Lavoisier for chemistry, Mendel for genetics or Seitz for materials science. I am sceptical about this suggestion, influential though such minds were, as a trawl through history shows that several streams founder to found a discipline, and the multiple founders are not coeval.

There are always those who turn down their thumbs at aspirant disciplines. The most celebrated instance is the remark attributed to Ernest Rutherford: “There’s physics, and there’s stamp collecting.” The proper response to this piece of hubris is to note that in Britain there’s a Royal Philatelic Society but an unroyal Institute of Physics. ■

Robert W. Cahn is in the Department of Materials Science and Metallurgy, University of Cambridge, Pembroke Street, Cambridge CB2 3QZ, UK.

FURTHER READING

Ziman, J. *Science Studies* 9, 67 (1996).
Ziman, J. *Real Science* (Cambridge Univ. Press, 2000).
Cahn, R. W. *The Coming of Materials Science* (Pergamon, Elsevier Science, 2001).



Turning over a new leaf

Paul Kenrick

A model based on biophysical principles of plant physiology, and drawing on fossil and environmental data, indicates that the origin of leaves was triggered by falling levels of atmospheric CO₂.

The early evolution of plant life on land was as much a conquest of the air as a subjugation of soil and rock. Leaves were one of the first innovations, and they have been such a success that it is difficult to imagine a world without them. Yet there is an enormous gulf of time between the origin of land plants and the evolution of leaves with a substantial lamina (flat blade) — for at least the first 40 million years of their existence, land plants were leafless or had only small spine-like appendages.

On page 352 of this issue, Beerling *et al.*¹ propose a novel explanation for this long delay. They use a model based on biophysical principles worked out from living plants, involving details of water use, and heat and gas exchange, and draw on anatomical and environmental data from the fossil record. They conclude that leaves with a broad lamina evolved in response to a massive drop in atmospheric CO₂ during the Devonian period, some 410–363 million years ago.

Plants with vascular systems — that is, with water- and food-conducting systems — have leaves of two main types. The microphyll is usually a short spine-like leaf and is characteristic of clubmosses; the megaphyll is generally bigger, and has a substantial lamina and a complex pattern of veins like the leaves of ferns and flowering plants². Almost all living plants have megaphylls or their derivatives, and it is this leaf type that principally concerns Beerling *et al.*

The gap between the earliest fossil evidence of vascular plants and the advent of megaphylls is puzzling. Megaphylls are ubiquitous today, and their photosynthetic proficiency makes their evolution seem inevitable. The structural framework necessary to assemble the primitive leaf (meristem, vasculature, cuticle and epidermis) was in place by the Early Devonian, some 410 million years ago. The fossil record shows that megaphylls evolved from the photosynthetic branching systems of early vascular plants that became flattened and later webbed to produce a broad lamina in several groups by the end of the Devonian³ (Fig. 1). Furthermore, the transformation of branch to leaf does not seem to require any major innovations in body plan. In the absence of obvious structural limitations, Beerling *et al.* turn to the environment for an explanation.

Certain aspects of plant structure change in response to fluctuations in atmospheric

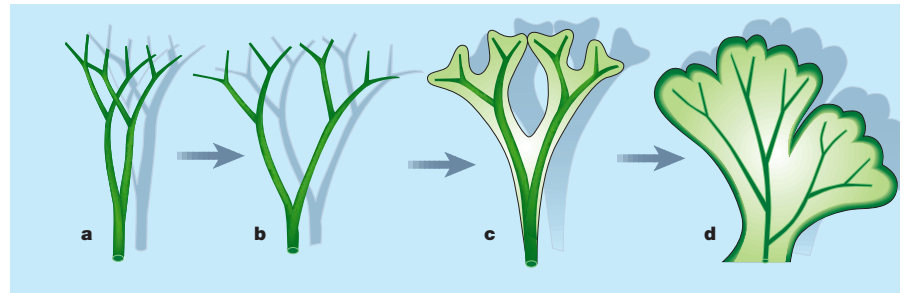


Figure 1 The advent of megaphylls — leaves with a broad lamina (flat blade), as opposed to the spine-like photosynthetic appendages of clubmosses. Fossil evidence shows that megaphylls evolved from simple, leafless photosynthetic branching systems in early land plants (a, b) to dissected (c) and laminate (d) leaves over a period of at least 40 million years. From Beerling and colleagues' biophysical model¹, which simulates gas exchange rates and energy budgets, it seems that these transformations were governed by falling concentrations of atmospheric CO₂ during the Devonian period, some 410–363 million years ago. In the high-CO₂ atmosphere that prevailed during the early phases of plant life on land, leaves with a broad lamina (c, d) and a low density of stomata would have been prone to overheating. Hence the absence of leaves with substantial blades during this time. Only with falling CO₂ in the atmosphere could megaphylls make their appearance, with their photosynthetic efficiency coming into its own.

gases, particularly CO₂. The density of stomata — the microscopic epidermal pores that regulate gas exchange and water transport — is one such variable⁴. Plants in a high-CO₂ environment need fewer stomata to supply their photosynthetic needs. But reduced stomatal density comes at a price: fewer stomata means that, at high light intensities, leaves with a broad lamina are more likely to suffer heat stress because less heat is lost through evaporation.

One response by plants to heat stress is to alter leaf shape, specifically the amount of dissection — the degree to which the margin of the leaf is lobed. Highly dissected leaves have deep, finger-like lobing and a greater cooling capacity than leaves with an unlobed blade. So, other things being equal, stomatal density decreases and leaf dissection increases with increasing levels of CO₂ (ref. 5). These variable traits can be seen in fossils, providing us with a kind of palaeontological barometer. Beerling *et al.* turn this approach on its head by using atmospheric CO₂ concentration to explain the development of megaphylls from photosynthetic branching systems, which are the functional equivalent of highly dissected leaves.

The evolution of megaphylls coincides with a drop of around 90% in atmospheric CO₂ concentration⁶. Working backwards, Beerling *et al.* use computer models to simulate gas exchange rates and energy budgets to

assess the likely costs and benefits of evolving leaves with a broad lamina under the high-CO₂ conditions that prevailed during the early phases of life on land. The model shows that a hypothetical megaphyll would overheat in such an atmosphere, reaching lethal temperature at low latitude and near-lethal temperature at high latitude. Furthermore, the rates of photosynthesis in branching systems exceed those of modelled high-latitude leaves, so having megaphylls at this early stage would not have been advantageous. In short, megaphylls would have made life too hot for the earliest plants.

The results of modelling systems that are so remote from the modern world have to be treated with scepticism. How can one be confident in a result for which so many parameters have to be estimated? This is a genuine concern, but Beerling and colleagues' calculations are based on independent biochemical, palaeobotanical and geochemical evidence, and seem reasonable. Enzyme kinetics can be estimated from living species; remarkable details of stem anatomy and stomatal distribution can be seen in the fossil record⁷; and the broad outline of atmospheric evolution in the Palaeozoic, although not written in stone, has withstood some independent tests⁸.

So there are good reasons to take Beerling and colleagues' proposal seriously. But it does require further scrutiny. For example, leaves

with a broad lamina would be expected to develop first at high latitudes, because here plants intercept less solar energy and are less likely to overheat. And one would expect high stomatal densities in early megaphylls because this would contribute to leaf cooling. Also, if fossils with megaphylls were found that dated to before around 386 million years ago, this would undermine the correlation between decreasing CO₂ and leaf lamination. Fossils with apparent leaf-like structures (protomegaphylls), such as *Eophyllophyton bellum* from 390-million-year-old deposits in China⁹, are of interest in this context.

Finally, if Beerling *et al.* are right, and leaf evolution was driven by a large fall in atmospheric CO₂, what then drove that fall in CO₂? Remarkably, the answer appears to be plants. The evolution of a land flora, and in particular the advent of large plants, had a major impact on the carbon cycle by drawing CO₂ out of the atmosphere and locking it up in trunks and leaves. Perhaps even more signifi-

cant, though, was the evolution of roots. The physical and chemical effects of root systems on rocks and soils increase rates of weathering, which is thought to have been responsible for removing enormous quantities of CO₂ from the Devonian atmosphere¹⁰. So roots may have played a key role in the evolution of leaves by way of the carbon cycle. ■

Paul Kenrick is in the Department of Palaeontology, The Natural History Museum, London SW7 5BD, UK.

e-mail: p.kenrick@nhm.ac.uk

1. Beerling, D. J., Osborne, C. P. & Chaloner, W. G. *Nature* **410**, 352–354 (2001).
2. Stewart, W. N. & Rothwell, G. W. *Paleobotany and the Evolution of Plants* (Cambridge Univ. Press, 1993).
3. Kenrick, P. & Crane, P. R. *Nature* **389**, 33–39 (1997).
4. McElwain, J. C. & Chaloner, W. G. *Ann. Bot.* **76**, 389–395 (1995).
5. McElwain, J. C., Beerling, D. J. & Woodward, F. I. *Science* **285**, 1386–1390 (1999).
6. Berner, R. A. *Am. J. Sci.* **294**, 56–91 (1994).
7. Edwards, D. *Phil. Trans. R. Soc. Lond. B* **353**, 141–157 (1998).
8. Mora, C. I., Driese, S. G. & Colarusso, L. A. *Science* **271**, 1105–1107 (1996).
9. Hao, S.-G. & Beck, C. B. *Palaeontographica B* **230**, 27–41 (1993).
10. Berner, R. A. *Phil. Trans. R. Soc. Lond. B* **353**, 75–82 (1998).

Astronomy

Brown dwarf is a radio star

Arnold O. Benz

Brown dwarfs are sometimes referred to as failed stars because they emit extremely little radiation. Astronomers have now found a brown dwarf that 'whistles' strongly at radio wavelengths.

Brown dwarfs are hard to find. They are small, cool objects, somewhere in size between the smallest stars and largest planets, and give off a dim, reddish optical glow. Their existence was confirmed only five years ago¹. Perhaps they would have been spotted earlier if they all emitted strong radio waves like brown dwarf LP944-20, described on page 338 of this issue by Berger and colleagues². This nearby brown dwarf (some 15 light years away) appears to be emitting a constant background of radio waves at a wavelength of 3.6 cm. These occasionally flare up for a few hours, growing in strength by more than an order of magnitude.

This is the first time that radio emission has been recorded from a bona fide brown dwarf. These objects are too small to burn hydrogen so they do not shine like regular stars, but they differ from giant planets because they can fuse deuterium in their centres. A brown dwarf has to be at least 12 times the mass of Jupiter to be able to burn some deuterium. This can only happen when it is young, in the first 10 million years or so. But at the ripe old age of 500 million years, LP944-20 is far beyond that stage and has no such heat source in its centre. So what might be producing its unusually strong radio emission?

Some of the smallest and coolest stars, like rapidly rotating type M and K stars, are known to be strong radio emitters at a

wavelength of 3.6 cm. A gaseous planet like Jupiter, on the other hand, emits appreciable radio waves only at wavelengths of ten metres and longer. So as far as radio emission is concerned, LP944-20 behaves more like a star than a planet.

In 1999, an X-ray flare was detected from LP944-20 for the first time³. To date, this is the only brown dwarf from which both X-ray and radio flares have been recorded. Active cool stars have a clear linear relationship between

their background radio and X-ray emission⁴ (Fig. 1). The relationship also holds for flares over many orders of magnitude⁵, from so-called nanoflares on the Sun to stellar superflares. To understand this relationship we have to know a little bit about how both types of radiation are created.

Cool stars have a surface temperature of just a few thousand degrees, but are surrounded by a much hotter atmosphere of ionized gases (plasma), known as the corona. Despite plasma temperatures reaching a few million degrees, the corona's low density means it yields relatively little thermal radiation. Nonetheless, the corona is responsible for most of the radio and X-ray emissions from these cool stars.

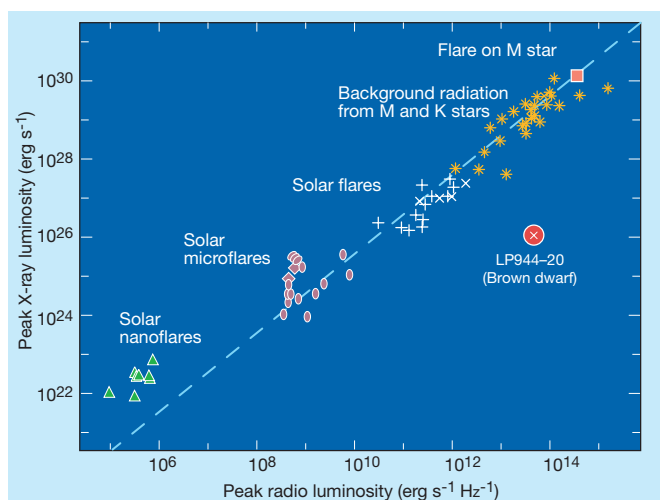
The hot gases in the corona emit thermal radiation as X-rays, which reach a peak during flares. Radio waves, on the other hand, are not produced by thermal radiation but through interactions between electrons and magnetic fields in the coronal plasma. The radio waves come from synchrotron radiation, which is emitted when energetic electrons spiral in the strong magnetic fields in the corona. By following the magnetic field lines, the electrons eventually hit the cool dense layer below the corona, where they heat it up to X-ray-emitting temperatures, resulting in flares.

The conditions for electron acceleration, propagation and energy deposition appear to be quite similar for various types of stars, resulting in the general relationship seen in Fig. 1. But deviations from this expected behaviour have been reported for very young stars⁶, and for some nanoflares deep in the Sun's corona⁷. In both cases, the radio emission was more than an order of magnitude lower than expected.

The observed radio emission of LP944-20 was more than three orders of magnitude stronger than expected, which is why it was detected. What makes this brown dwarf so different from cool stars? The observed rela-

Figure 1 Peak fluxes of X-ray and radio emissions from solar and stellar flares at centimetre wavelengths⁶. The relationship between radio and X-ray emission is consistent across a wide range of stars, from the coolest M and K types, to ordinary G stars like our Sun. The background emission from cool M and K stars that are more active than the Sun may just be the product of many small flares. Berger *et al.*¹ have

now discovered surprisingly strong radio emission from a brown dwarf, LP944-20, which is smaller than the coolest stars, so is not expected to shine brightly at any wavelength.



tionship between radio and X-ray emissions is an empirical one, and the underlying physical causes may be quite different. For instance, if energetic electrons live a lot longer before they lose their energy, the total synchrotron radiation of a source is accordingly stronger.

Alternatively, electrons may emit radio waves through a completely different mechanism in brown dwarfs. If energetic electrons are trapped by a magnetic field, they can generate radio waves by the very efficient 'cyclotron maser' process, in which bunches of electrons emit coherent waves. Or plasma waves can be generated that are transferred into propagating radio waves. Such mechanisms can produce radio waves far stronger than would be possible by synchrotron radiation alone. A maser process is generally assumed to be the cause of Jupiter's 10-metre radio waves, and coherent radio bursts from the Sun and other stars at decimetre and centimetre wavelengths are well known.

Berger *et al.*² give several reasons why they believe that the radio emission of LP944-20 originates from the synchrotron mechanism. They suggest that LP944-20 has an unusually weak magnetic field compared to other brown dwarfs, and even to Jupiter, perhaps allowing extended synchrotron emission from long-lived electrons. This model can be confirmed by simultaneous observations across all wavelengths and

by higher-resolution observations at radio wavelengths.

If the magnetic field of LP944-20 is as weak as Berger *et al.* suggest, then an enormous number of highly energetic electrons would be required for the synchrotron process to work. The creation of so many electrons moving near the speed of light would require an efficient and productive acceleration mechanism, as well as extraordinarily good trapping of particles within the corona. But for the most plausible alternative, maser emission, to contribute to the radio signal from LP944-20, the magnetic field would have to be more than 100 times higher than that inferred by Berger and colleagues.

Both types of emission mechanism far exceed Jupiter's capacities and would be remarkable even for a brown dwarf. The whistling brown dwarf is a new mystery for astronomers to puzzle over, and may indicate that these dusky little objects have more surprises in store.

Arnold O. Benz is at the Institute of Astronomy, ETH-Zentrum, CH-8092 Zurich, Switzerland.
e-mail: benz@astro.phys.ethz.ch

1. Nakajima, T. *et al.* *Nature* **378**, 463–465 (1995).
2. Berger, E. *et al.* *Nature* **410**, 338–340 (2001).
3. Rutledge, R. E., Basri, G., Martin, E. L. & Bildsten, L. *Astrophys. J.* **538**, L141–L144 (2000).
4. Guedel, M. & Benz, A. O. *Astrophys. J.* **405**, L63–L66 (1993).
5. Benz, A. O. & Guedel, M. *Astron. Astrophys.* **285**, 621–630 (1994).
6. Smith, K. W., Guedel, M. & Benz, A. O. *Astron. Astrophys.* **349**, 475–484 (1999).
7. Krucker, S. & Benz, A. O. *Solar Phys.* **191**, 341–358 (2000).

Molecular clocks

A vivid loop of light

Matthew P. Pando and Paolo Sassone-Corsi

Studies of biological molecular clocks have benefited greatly from a unicellular fungus. The latest discovery is of a fungal protein required in resetting the clock in response to light.

Throughout the course of evolution, there has been a selective advantage for most organisms to regulate their behaviour and physiology according to a daily (circadian) cycle. The internal molecular clocks that have evolved, which keep us all on a regular schedule, operate with roughly a 24-hour period. One seeming limitation that all biological clocks have is that they constantly need minor adjustments, termed 'entrainment'. In reality, however, this is not a limitation but rather a daily fine-tuning that ensures that all the biological rhythms controlled by circadian clocks are in phase with the environment. One aspect of the environment that can always be counted on is the sunrise, and this is no doubt why light has become the major signal that keeps biological clocks in sync with the outside world¹.

The mechanisms that regulate the clocks' response to light and their ability to adjust themselves to environmental light cycles

are only just being deciphered. One such mechanism is now described in *Cell* by Heintzen and colleagues² and involves a protein called VIVID, which is important in controlling responses to light in the fungus *Neurospora crassa*. The authors show that VIVID is a clock-associated self-regulator, which inhibits itself by means of a negative-feedback loop that also controls the link between light input and the central circadian clock in *Neurospora*. This is one of the first examples of an autoregulatory loop that links the regulation of the central clock directly to the clock's input and output pathways.

Heintzen *et al.*² show that the expression of the *vivid* gene, and production of the VIVID protein, is rapidly (but temporarily) induced in response to light. Like many components of the central clock mechanism, VIVID contains a so-called PAS domain — a region that allows it to interact with other proteins. The authors suggest that VIVID and

a component of the central *Neurospora* clock, a protein called WC-1, interact through their respective PAS domains. It was already known that WC-1, in response to light, forms a complex with another protein, WC-2, and that together they activate the expression of genes that are themselves either part of the clock or part of the clock's output. But when VIVID forms a transient complex with WC-1, this activation of gene expression is blocked. It appears that the light-induced expression of VIVID itself is directly regulated by the WC-1–WC-2 complex (Fig. 1). So in effect, by blocking the activity of this complex, VIVID leads to its own downregulation — this is a negative-feedback loop. And, by functioning as a clock-associated negative regulator of the WC-1–WC-2 complex, VIVID also regulates other light-responsive, clock-controlled genes.

The VIVID protein also has another function — in 'gating', a term used to describe the mechanisms that allow the central circadian clock to be unresponsive to light during a certain period of the day–night cycle³. This 'dead zone' typically falls during the middle of the day. Heintzen *et al.* show that, in the absence of VIVID, the gated response to light is still present but altered. In addition, the amplitude of light-induced gene expression

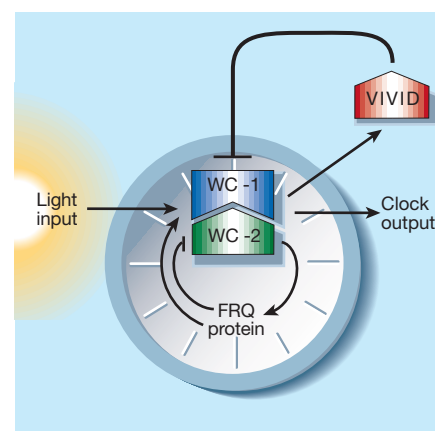


Figure 1 Controlling the effect of light on the molecular clock in *Neurospora crassa*. The central clock comprises a self-regulating loop of the WC-1, WC-2 and FREQUENCY (FRQ) messenger RNAs and proteins. The output of the clock consists of the oscillating expression of genes that are involved in aspects of behaviour and physiology that vary according to a daily cycle. The clock keeps in tune with the environment in part by receiving light input. As shown by Heintzen *et al.*², the VIVID protein is essential for the clock's response to light. The expression of VIVID is induced in response to light by the WC-1–WC-2 complex; the VIVID protein then binds to WC-1 and inhibits the induction of expression of FRQ and clock-output genes by WC-1–WC-2. VIVID is also required for resetting the clock in response to light, and in 'gating' — the portion of the day when the clock becomes unresponsive to light.

is greater, and the gene transcripts have longer half-lives than normal. It seems that although VIVID is not solely responsible for generating a gated light response, it is involved in maintaining the accuracy of gating in *Neurospora*. Finally, Heintzen *et al.* show that VIVID is also needed for resetting the *Neurospora* clock in response to light; the protein controls what type of effect a light signal has on the central clock.

Neurospora has once again proven to be an invaluable model for advancing our knowledge of the circadian system. In this latest example, Heintzen *et al.* have provided strong evidence for a feedback loop that receives information from the central clock so as to regulate the effects of light input to the clock. Yet, despite its importance to our understanding of circadian biology⁴, *Neurospora* is still a unicellular fungus; the control of circadian rhythms in multicellular animals is more complicated. Clocks that are independent of the central pacemaker are found in a variety of tissues in multicellular organisms^{5,6}. In semi-transparent creatures such as fruitflies and zebrafish, these 'peripheral' clocks respond directly to light^{7,8}. For these animals, it is reasonable to think that a regulatory loop similar to that described for VIVID in *Neurospora* could exist, and might be responsible for regulating responses to light and clock resetting in both central and peripheral clocks.

On the other hand, opaque animals, such as mammals, rely on the detection of light signals by anatomical structures that have evolved to do so, such as the retina and pineal gland. These structures must process the light signal so as to adjust not only the peripheral clocks but also the master regulator — the suprachiasmatic nucleus, located in the brain⁹. Perhaps, in the mammalian retina, there is an autoregulatory VIVID loop that is similar to that found in *Neurospora*, regulating mammalian responses to light and modulating the neuronal signal transmitted from the retina to the suprachiasmatic nucleus.

One could imagine that there is a trickle-down effect, and that resetting the mammalian clocks involves an initial modulation of light input by a retina-based VIVID loop. But, as with all respectable governments, there are always checks and balances. It would make sense for three types of VIVID loop to operate in mammals. A primary loop would respond to and regulate light signals in the retina; a secondary loop would modulate neuronal signals sent from the retina to the suprachiasmatic nucleus; and finally tertiary loops would respond to master-clock-controlled chemical signals responsible for resetting the peripheral pacemakers.

But before we blind ourselves with too many VIVID loops, it will be important to find out whether or not equivalents of the VIVID protein even exist in higher organisms, or whether VIVID's functions have

been incorporated into other proteins. If mammalian equivalents do exist, they may have the same, extra or totally new functions in a given circadian system. Take cryptochromes, for example. These pigments are involved in light detection in the plant *Arabidopsis thaliana* and in fruitflies^{10,11}, but in mammals they no longer have a major role in detecting light, instead functioning as components of the central clock¹². As important as Heintzen *et al.*'s findings are, only time will tell if other organisms have a similarly regulated clock.

Matthew P. Pando and Paolo Sassone-Corsi are at the Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS-INSERM-ULP, 1 rue Laurent

Fries, 67404 Illkirch-Strasbourg, France.

e-mail: paolosc@igbmc.u-strasbg.fr

1. Dunlap, J. C. *Cell* **96**, 271–290 (1999).
2. Heintzen, C., Loros, J. J. & Dunlap, J. C. *Cell* **104**, 453–464 (2001).
3. Daan, S. & Pittendrigh, C. S. *J. Comp. Physiol.* **106**, 253–266 (1976).
4. Bell-Pederson, D. *Fungal Genet. Biol.* **29**, 1–18 (2000).
5. Yamazaki, S. *et al. Science* **288**, 682–685 (2000).
6. Balsalobre, A., Damiola, F. & Schibler, U. *Cell* **93**, 929–937 (1998).
7. Plautz, J. D., Kaneko, M., Hall, J. C. & Kay, S. A. *Science* **278**, 1632–1635 (1997).
8. Whitmore, D., Foulkes, N. S. & Sassone-Corsi, P. *Nature* **404**, 87–91 (2000).
9. Cermakian, N. & Sassone-Corsi, P. *Nature Rev. Mol. Cell Biol.* **1**, 59–67 (2000).
10. Somers, D. E., Devlin, P. F. & Kay, S. A. *Science* **282**, 1488–1490 (1998).
11. Stanewsky, R. *et al. Cell* **95**, 669–679 (1998).
12. van der Horst, G. T. J. *et al. Nature* **398**, 627–630 (1999).

Solar physics

Sizing up the Sun

Douglas Gough

Variations in sunspot activity affect the weather on Earth, but the underlying mechanisms are unknown. New measurements of the Sun's radius and luminosity may help pinpoint the source of the energy responsible.

Minor physical changes in the Sun often lead to extreme solar magnetic activity, which can affect the Earth — disrupting radio communications and influencing the weather, for instance. Although an understanding of the mechanisms that cause the 11-year variation in solar magnetic activity (also known as the solar cycle) is of considerable interest, it is currently beyond our reach. Fortunately, increasingly accurate and detailed observations of the Sun seem to be drawing us closer. Emilio *et al.*¹ report the results of sensitive satellite observations in the *Astrophysical Journal*: they find evidence that the energy responsible for variations in the Sun's radius and luminosity does not come from the inner depths of the Sun but rather the outer layers (Fig. 1). This observation is certainly not the first claimed detection of a small variation in the Sun's radius, but it may be the first to survive the test of time.

The Sun's magnetic activity varies cyclically — most visibly with the number of sunspots — over an 11-year time period. We are currently experiencing a peak in the solar cycle and therefore in the number of sunspots. It is generally believed that the underlying cause of this cycle is the interaction between the Sun's rotation and a 'dynamo' responsible for the Sun's magnetism. In this model, the dynamo effect creates a magnetic field from the electrical currents caused by convection and large-scale shearing motion within the Sun. But the Sun is a globe of hot gas, so it does not rotate like a solid; the outer regions tend to rotate faster near the equator and slower near the poles, which results in stronger magnetic fields bursting through the photosphere — the visible surface of the

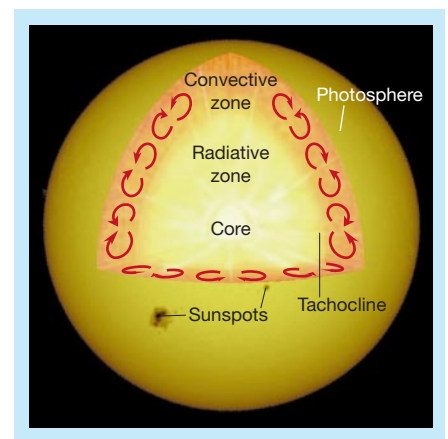


Figure 1 The interior structure of the Sun.

Energy is generated by nuclear fusion in the core, which passes through the radiative zone as light and heat. The energy is transferred to the Sun's surface (photosphere) by the layer of convective gas. Emilio *et al.*¹ have found that the Sun's radius varies over time with the number of sunspots (cool darker regions with extreme magnetic fields) on its surface.

Sun. These magnetic fields inhibit the convective transport of heat, permitting material to cool at the surface, and producing visibly darker regions known as sunspots.

None of the dynamo models explains the small, observed variation in the Sun's luminosity, which also follows the sunspot cycle. These small changes in luminosity (no bigger than 0.1%) come from the release of stored energy somewhere within the Sun, which also affects the size of the Sun. So by studying the tiny changes in the radius we might learn something about the source of this extra

energy, and ultimately the nature of the process that causes the luminosity change.

Since the late 1970s, solar physicists have tried to calculate or measure the ratio (W) of the relative radius and luminosity changes over a solar cycle². In the early days, a common motivation for calculating W was to provide a means for inferring luminosity changes from measurements of radius, and hence to study the influence of luminosity on the Earth's climate. It was not anticipated that reliable measurements of luminosity changes would precede reliable measurements of radius changes. Observations from space have now shown that solar irradiance, which is believed to be closely linked to luminosity changes, oscillates with an amplitude of 0.05%, the maxima occurring at times of maximum magnetic activity in the solar cycle³. So now the interest is in measuring radius changes to obtain W , in order to set a constraint on where in the Sun the stored energy is modulated to produce the variations in both radius and luminosity.

Many theoretical estimates of W have been published. Almost all of them are based on numerical models of the Sun's response to a sudden change in convection — the transfer to the Sun's surface of heat generated in the core. The idea is that the Sun's magnetism affects heat transport in the convective layer in a way that influences both luminosity and radius. Despite many contradictory results produced by early calculations, it was found⁴ that on the whole W increases with increasing depth of the mechanism that causes the variation in luminosity and radius. For example, a calculation that modifies the effectiveness of convective heat transfer, which influences only the very outer layers of the Sun, yields $W \approx 0.002$ (refs 5,6), whereas studies that locate the mechanism at the base of the convective zone or in the core have $W \approx 0.2$ and $W \approx 0.5$, respectively⁴.

There is a long history of attempts to measure the Sun's radius from the ground. The efforts have been plagued by the distorting effects of the Earth's atmosphere, and there is enormous disparity among the results. One of the more reliable investigations, perhaps, is the recently published measure of the photospheric radius by Brown and Christensen-Dalsgaard⁷, from which one can infer that $W \leq 0.1$. This upper limit is consistent with most of the theoretical values but it does eliminate the Sun's core as the location of the extra stored energy, because it is less than 0.5.

To avoid atmospheric distortion, the measurements of Emilio *et al.*¹ were made from space, using the Michelson Doppler Imager⁸ on the Solar and Heliospheric Observatory (SOHO). There was, however, some contamination due to variation in the temperature of the instrument as the distance of SOHO from the Sun varied, which induced a slight change in the focal length

of the optical system. The contamination has a period of one year, the orbital period of SOHO, and so could easily be removed from the data. The authors found that the remaining radius change increases approximately in a linear way with the number of sunspots. The satellite observations were made over three years, a period which does not extend over one solar cycle. But by extrapolating the data over existing sunspot records, the value $W \approx 0.04$ can be inferred. (This value differs from that quoted by Emilio *et al.*, who erroneously doubled the value of the observed irradiance change.) The authors recognize that there are uncertainties in the interpretation of their data, and suggest that their value might actually be an upper limit.

What can one conclude from the results? If the upper limit is correct, the principal cause of the variation is not likely to lie very deep inside the Sun. Even if, as is popularly believed, the tachocline — a thin layer of rotational shear beneath the base of the convection zone — were to be playing a major role in the dynamics of the cycle, the

associated luminosity variations must be induced further away from the core, since $W < 0.1$. But one might eventually deduce more. If it were to transpire that 0.04 is actually a good estimate of, rather than an upper limit to, W , because it is much greater than 0.002 it would imply that the luminosity change is not simply a result of a change in the effectiveness of convection at the surface; the principal cause would have to be more deeply seated. We wait with keen anticipation to see whether that is so. ■

Douglas Gough is in the Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, UK.
e-mail: douglas@ast.cam.ac.uk

1. Emilio, M., Kuhn, J. R., Bush, R. I. & Scherrer, P. H. *Astrophys. J.* **543**, 1007–1010 (2000).
2. Sofia, S. *et al. Science* **204**, 1306–1308 (1979).
3. Fröhlich, C. *Space Sci. Rev.* **94**, 15–24 (2000).
4. Gough, D. O. in *NASA Conf. Publ.* 2191 (ed. Sofia, S.) 185–206 (NASA, Washington DC, 1981).
5. Däppen, W. *Astron. Astrophys.* **124**, 11–23 (1983).
6. Balmforth, N. J. *et al. Mon. Not. R. Astron. Soc.* **278**, 437–448 (1996).
7. Brown, T. M. & Christensen-Dalsgaard, J. *Astrophys. J. Lett.* **500**, L195–L198 (1998).
8. Scherrer, P. H. *et al. Solar Phys.* **162**, 129–188 (1995).

Neurobiology

New memories from new neurons

Jeffrey D. Macklis

Neurons are born throughout adulthood in specific regions of mammalian brains. It now turns out that — in rats at least — these new neurons are essential for the formation of one type of memory.

There is a long-standing dogma that the brains of adult mammals do not generate new nerve cells. Yet data to the contrary have existed for over 35 years^{1–3}, and more recent results have brought general acceptance that neurons are indeed born in the adult mammalian brain — but only in two regions. These are the olfactory bulb, which is involved in sensing smell^{4,5}, and the highly complex hippocampus, involved in memory⁶. Neurons are born in these regions even in 'higher' adult mammals, such as monkeys and humans^{7–9}. But do the new neurons have a crucial role, or are they subservient to permanent members of the neuronal circuitry? On page 372 of this issue, Shors and colleagues¹⁰ address this question. They present evidence that the addition of new neurons to the hippocampus of adult rats is essential for at least one form of memory, that concerning the timing of learned responses, and temporal relationships between events (Fig. 1).

Much work has gone into uncovering the effects of learning, hormones, seizures, physical activity and the environment on the birth of new neurons (neurogenesis) in the hippocampus of adult rodents¹¹. Neurogenesis and the subsequent survival of new neurons appear to be modulated by activities

that could be interpreted as involving the hippocampus. For example, more inquisitive physical activity appears to result in more new neurons being recruited into the hippocampus. But why? Do the new neurons actually contribute to the function of the adult brain? The level of neurogenesis appears to be decreased by stress, an uninteresting environment and perhaps depression. Again, why? And does this reduction have harmful effects on memory and behaviour?

The answers to these questions will be important not only in understanding the role of neurogenesis in the function of the hippocampus, but also in the field of brain repair. If the development of immature neurons and precursor cells could be controlled, those cells might replace neurons that are dead or dying, as a result of neurodegenerative disease, stroke or spinal-cord injury, for example. But the obstacles to achieving such precise control over neuronal fate are immense. Equally important, if these approaches are to be successful, the newly incorporated neurons must be able to function in the existing neuronal circuitry. But can they?

Studies of other species, particularly songbirds, are starting to provide some answers. Throughout adulthood, new neu-

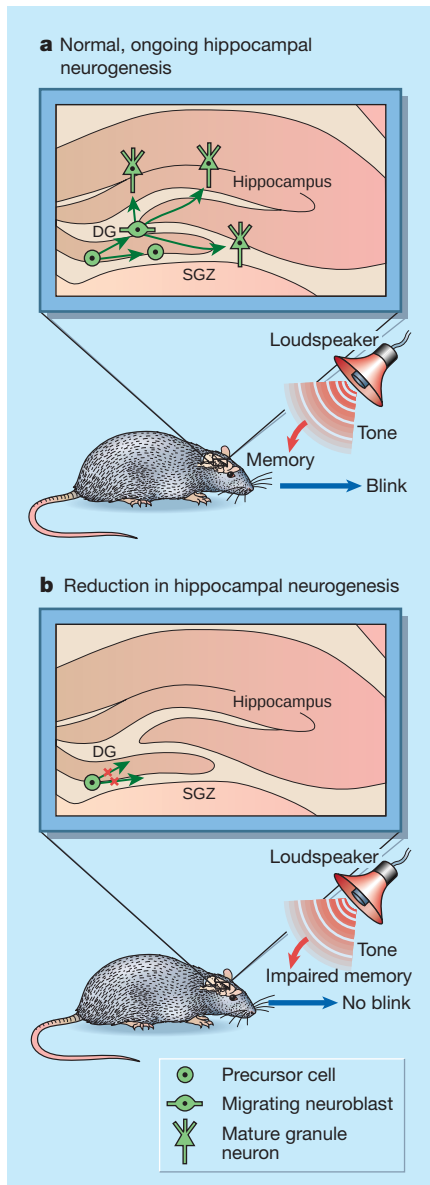


Figure 1 New nerve cells that are born in the hippocampus of the adult mammalian brain are crucial in certain types of memory formation. In the dentate gyrus (DG) of the hippocampus in adult mammals, new neurons appear to be born from the division of precursors in the subgranular zone (SGZ). Shors *et al.*¹⁰ investigate whether these neurons integrate themselves functionally into existing brain circuitry by studying a type of memory called 'trace conditioning', both under normal conditions (a) and following reduction of neurogenesis, the birth of new neurons (b). Rats were first trained to associate a neutral tone with a delayed, mildly unpleasant stimulus. a, Under normal conditions, rats later blinked in response to the neutral sound alone. b, Shors *et al.* reduced neurogenesis in the hippocampus by about 80% using a drug that kills proliferating cells. This impaired hippocampus-dependent trace-conditioning memory, suggesting that the addition of new neurons to the hippocampus of adult rats is essential for some types of memory related to the timing and temporal order of events.

rons are added to complex brain regions in many bird species^{12,13}. Correlative evidence suggests that new neurons are involved in song production and in memory, and recent results have revealed the function of experimentally recruited neurons¹⁴. Studies of rodents have shown that transplanted immature neurons and precursor cells can, to some degree, repair even complex, long-distance connections in the adult brain¹⁵. It is even possible to induce the birth of new neurons in the cerebral cortex of adult mice, a region where neurogenesis does not normally occur¹⁶. These neurons, too, can make long-distance connections. But the question remains as to whether the thousands of neurons born every day in the hippocampus of adult rats contribute to brain function. Shors *et al.*¹⁰ meet this question head-on.

The authors have studied a type of memory called 'trace conditioning'. Here, rats are first trained to associate a neutral (neither pleasant nor unpleasant) sound with a delayed, unpleasant stimulus — in this case,

gentle stimulation of the eyelid. The rats are then tested to see whether they react (by blinking) to the neutral sound alone (Fig. 1a), a sign that they are remembering the unpleasant stimulus. Unlike some other types of memory, trace conditioning depends on the hippocampus.

Shors *et al.* reduced neurogenesis in the hippocampus by using the drug methylazoxymethanol (MAM), which kills proliferating cells, and examined the effects on the trace-conditioning memory task. Any new cells could be identified because they were labelled with bromodeoxyuridine, an analogue of the DNA base thymidine that becomes incorporated into the new DNA formed during cell replication, and which can be detected by dye-labelled antibodies. In this way, the authors gauged the effectiveness of treatment with MAM. This drug also prevents cells other than neurons from proliferating. So Shors *et al.* included experiments to ensure that there was no nonspecific damage to existing hippocampal

neurons, and that the results were not due simply to stress or general illness in the rats.

The authors found that a roughly 80% reduction in the number of newborn neurons in the adult hippocampus impaired the hippocampus-dependent trace-conditioning memory (Fig. 1b), but had no effect on another, hippocampus-independent form of memory. Restoring normal levels of neurogenesis in the hippocampus, after the end of the treatment with MAM, led to the recovery of trace-conditioning memory. The implication is that the normal level of neurogenesis in the hippocampus of adult rats is required for some types of memory that are related to the timing and temporal order of events. By extension, it seems that the new neurons themselves are involved in forming new memories.

But the evidence that new hippocampal neurons have a functional role is, as yet, merely correlative. Some questions remain unanswered. For example, as hinted above, MAM is toxic to all dividing cells in the body, so it might affect other brain regions or make the rats stressed or sick. Stress has been shown to affect memory. If stress were induced by MAM, it might not affect all types of memory equally; the relatively difficult hippocampus-dependent task might be affected more than the easier, hippocampus-independent task used as a control. The authors chose a dose of MAM that they believe avoids these problems, and — by observing the rats grooming and by measuring steroid levels in the rats' blood — they found no evidence that the animals were sick or very stressed. But more subtle nonspecific effects cannot be entirely ruled out. Also, the observed effects appeared sooner than might have been predicted from previous results.

That said, Shors *et al.*¹⁰ have begun to answer the question of whether neurogenesis in the adult hippocampus contributes to memory function. The data suggest that proliferating cells in the adult brain are essential for the continual renewal of the hippocampal nerve circuits involved in certain types of temporal memory. The results also support the idea that it might, one day, be possible to add new, fully functional neurons into existing brain circuitry to treat diseases of the nervous system. ■

Jeffrey D. Macklis is in the Division of Neuroscience, Children's Hospital, and the Department of Neurology and Program in Neuroscience, Harvard Medical School, Enders 354, 320 Longwood Avenue, Boston, Massachusetts 02115, USA.
e-mail: jeffrey.macklis@tch.harvard.edu

- Altman, J. & Das, G. D. *J. Comp. Neurol.* **124**, 319–335 (1965).
- Altman, J. *J. Comp. Neurol.* **137**, 433–457 (1969).
- Kaplan, M. S. & Hinds, J. W. *Science* **197**, 1092–1094 (1977).
- Luskin, M. B. *Neuron* **11**, 173–189 (1993).
- Lois, C. & Alvarez-Buylla, A. *Proc. Natl Acad. Sci. USA* **90**, 2074–2077 (1993).
- Kuhn, H. G., Dickinson-Anson, H. & Gage, F. H. *J. Neurosci.* **16**, 2027–2033 (1996).
- Gould, E. *et al. Proc. Natl Acad. Sci. USA* **96**, 5263–5267 (1999).

8. Kornack, D. R. & Rakic, P. *Proc. Natl Acad. Sci. USA* **96**, 5768–5773 (1999).
9. Eriksson, P. S. *et al. Nature Med.* **4**, 1313–1317 (1998).
10. Shors, T. J. *et al. Nature* **410**, 372–376 (2001).
11. Gage, F. H. *Science* **287**, 1433–1438 (2000).
12. Goldman, S. A. & Nottebohm, F. *Proc. Natl Acad. Sci. USA* **80**, 2390–2394 (1983).

13. Barnea, A. & Nottebohm, F. *Proc. Natl Acad. Sci. USA* **91**, 11217–11221 (1994).
14. Scharff, C., Kirn, J. R., Grossman, M., Macklis, J. D. & Nottebohm, F. *Neuron* **25**, 481–492 (2000).
15. Shin, J. J. *et al. J. Neurosci.* **20**, 7404–7416 (2000).
16. Magavi, S. S., Leavitt, B. R. & Macklis, J. D. *Nature* **405**, 951–955 (2000).

Biogeochemistry

Oxygenating the atmosphere

Norman H. Sleep

Photosynthesis is the main source of oxygen in Earth's atmosphere. But it may have been geological activity that first allowed an oxygen-rich atmosphere to develop.

We usually give little thought to the air we breathe. Yet the Earth's atmosphere was not always as rich in oxygen as it is today. Oxygen now constitutes about 20% of the gas in the atmosphere, but before about 2,500 million years ago it was only a trace constituent. Writing in the new journal *Geochemistry, Geophysics, Geosystems*, Kump, Kasting and Barley¹ propose that changes in the deep interior of the Earth affected the composition of volcanic gases, and that this led to the rise in atmospheric oxygen levels at the Archaean–Proterozoic transition, 2,470 million to 2,450 million years ago. In turn, the increase in oxygen allowed complex life to evolve.

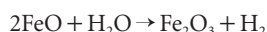
The Earth is unique among the planets in the Solar System in having an oxygenated atmosphere. The atmospheres of the other planets are anoxic because oxygen levels are kept comparatively low by an equilibrium system involving chemical processes — in, for instance, volcanic gases, and the planets' crusts and mantles. On the Earth, oxygen levels built up over geological time as a result of photosynthesis, which can be expressed as the reaction $\text{CO}_2 \rightarrow \text{C} + \text{O}_2$. Over the years, a vast amount of organic carbon has become locked up in sedimentary rocks, a small part of it as coal and oil, so preventing the reverse reaction to equilibrium that would create CO_2 and lower atmospheric levels of oxygen. Today, carbon burial is an inefficient process — only 0.1% of the carbon produced by photosynthesis ends up in sedimentary rocks, the rest being oxidized — which helps to keep today's oxygen levels stable.

But the advent of oxygen-producing photosynthesis cannot be the whole story. Photosynthesis existed long before the Archaean–Proterozoic transition and the well-documented rise in oxygen levels, as is evidenced by the existence of photosynthetic bacteria 3,500 million years ago. But it is possible that these bacteria were part of a global microbial ecology that used sulphate as the main oxygen carrier, and so would consume only trace amounts of oxygen.

Instead of looking at what processes

might produce oxygen, Kump *et al.*¹ consider the materials and reactions that consume it and might have kept its levels low. They outline a scheme in which the reduced, oxygen-consuming gases carbon monoxide (CO), hydrogen (H_2) and methane (CH_4) were vented from arc volcanoes on land and midocean ridges beneath the sea (Fig. 1, overleaf). These reducing gases emanating from the mantle were a major sink for the oxygen resulting from photosynthesis in the Archaean, each reacting with oxygen to produce CO_2 , water or both.

In addition, rocks containing a reduced form of iron — iron(II) or ferrous iron — were another sink for the oxygen produced by photosynthesis. These rocks were basalts that erupted on the ocean floor or land surface, and they sequestered the oxygen from water using the oxidation reaction that turned ferrous iron into its ferric — iron(III) — counterpart.



This reaction is key to Kump *et al.*'s hypothesis because it traps oxygen through oxidation of mantle rock, and regenerates a reducing gas.

Kump *et al.* believe that subsequent events in the Earth's mantle explain the transition to an oxygen-rich atmosphere. This transition began abruptly at the Archaean–Proterozoic boundary, which is thought to coincide with a change in tectonic conditions in the Earth. In the authors' model, the sudden change is a consequence of the poor mixing — stratification — in the mantle.

The idea is that cold, dense, oxidized slabs, containing high amounts of iron(III) oxide, sank rapidly from close to the surface of the Earth and settled near the boundary between the Earth's core and mantle. Eventually, most of the lower mantle was filled with this oxidized material. As today, the mantle was heated from below by the core and from within by radioactivity. Heating from the core meant that the base of the mantle became buoyant about 2,700 million



100 YEARS AGO

Sir Courtenay Boyle objects, in the March number of *Macmillan's Magazine*, to many words in common use in science. His objections are partly etymological and partly to the vagueness of connotation of the words. Pliocene, miocene and phonolite are incorrectly formed; and the first two, together with palaeozoic, mesozoic, kainozoic, jurassic and triassic are condemned because they are purely relative terms. Electron is objected to because there is sometimes a doubt whether it signifies a minute corpuscle having an electric charge or an electric charge without the corpuscle. Kion and autokion are suggested as preferable to the unsatisfactory words motor and the hybrid automotor.

From *Nature* 14 March 1901.

50 YEARS AGO

Mr. Ritchie Calder, the science editor of the *News Chronicle*, was sent by Unesco to North Africa and the Middle East to report on what is being done to reclaim the deserts. The strange aura of romance and the mystery that surrounds the desert attracted much attention to the assignment, which was reported (according to the publisher's note) in thirty-two countries... In Israel... and to a lesser extent in Iraq and Persia, desert reclamation is a matter of urgent practical politics. Those beings whom Mr. Calder calls "the scientists" and who correspond to the gods who wrought wonders in former ages are being called upon to shower their blessings on mankind. Iraq and Persia have in their oil great wealth, some of which can be diverted to restoring the productivity of the land by modern adaptation of ancient methods. Israel has no mineral wealth, and her greatest resource in conquering the desert is the ingenuity of her people, and particularly of "the scientists" who are not only planning great irrigation and hydroelectric schemes but also are working on the large-scale desalinization of salt water by ionic exchange and by distillation through nylon. Promising experiments are being made to use the opacity of flake nylon to light and heat to reduce evaporation from reservoirs. There seem to be possibilities that nylon may become a key material in the conservation of water in Israel, and schemes are being considered for the large-scale cultivation of the castor-oil plant as a source of raw material for nylon manufacture.

From *Nature* 17 March 1951.

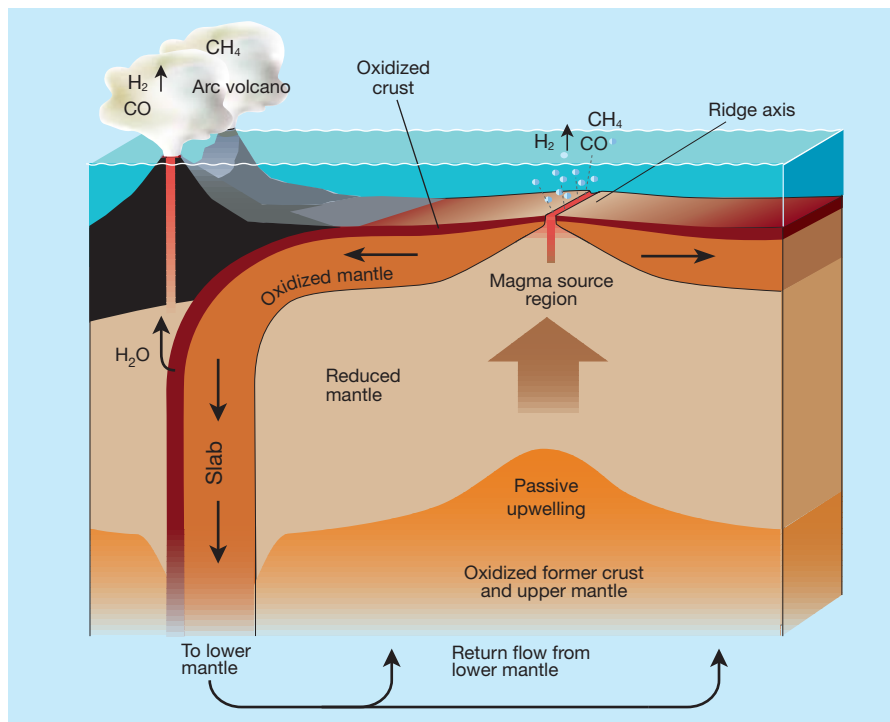


Figure 1 Events near the end of the Archaean, around 2,500 million years ago, which led from an oxygen-depleted to an oxygen-rich atmosphere. Volcanic gases emanating at arc volcanoes and midocean ridges from a reduced mantle were themselves reduced and hence were an oxygen sink. In the crust and mantle, the oxidation of iron(II) oxide to iron(III) oxide also helped to keep oxygen levels low. According to the course of events outlined by Kump *et al.*¹, slabs of cold, dense, oxidized upper mantle and crust sank by subduction deep into the lower mantle. Heating from the core led to the rise of this oxidized material. Eventually, by overturn of the lower mantle into the upper mantle, or by filling the mantle completely, this oxidized material became the source of gases released at arc volcanoes and midocean ridges. As a result, the gases were insufficiently reduced to be a significant sink for oxygen — hence the rise in oxygen concentrations in the atmosphere.

years ago, causing plumes of oxidized rocks to ascend towards the surface. Kump *et al.* suggest that the Archaean ended when the whole oxidized lower mantle became buoyantly unstable and overturned into the upper mantle, so that oxidized material replaced the reduced iron(II).

An alternative explanation, which gives the same outcome as far as oxygenation of the atmosphere and chemical mass balances are concerned, is that oxidized slabs filled the

entire mantle by the end of the Archaean, replacing the reduced upper mantle that had never existed in slab form².

These events made the mantle source regions for magmas less reducing, which in turn had the same effect on the volcanic gases emanating from them. Eventually, the mantle became oxidized to the extent that volcanoes and midocean ridges vented gases that were an ineffective oxygen sink. Oxygen levels could then build up in the air, and the

levels of H₂, CH₄ and CO were brought down to their present status as trace constituents of the atmosphere.

The basic geometry of the scheme proposed by Kump *et al.* works out; a plausible amount of oxidized slab material could have filled the mantle in the given time. It is also the case that the Earth's mantle is more oxidized than the material from which it formed. The success of Kump and colleagues¹ hypothesis hinges on the time that the Earth's mantle became oxidized. At present, data from Archaean rocks come only from volcanic rocks generated by mantle plumes, which — as expected from the authors' reasoning — are oxidized. What are needed are data from Archaean volcanic rocks from an upper mantle source (such as the basalts erupted at mid-oceanic ridges), which according to Kump *et al.* should be reduced. In other words, their hypothesis is testable with data that are mostly independent of the surface carbon–oxygen cycle.

As well as looking at the rise of photosynthesis, traditional approaches to the increase of oxygen in the atmosphere have concentrated on surface biogeochemical processes. Carbon burial has been the subject of especial attention because, if free to react, the amount of carbon now sequestered in sediments would be a huge oxygen sink³. But although carbon burial is a necessary condition for maintaining a high-oxygen atmosphere, it may not have been a sufficient one for generating it in the first place. Controls on the process — which in part depends on availability of the nutrient phosphorus from rock weathering⁴ — are not well understood.

Finally, it is also plausibly the case that weathering of volcanic rocks at the surface was a controlling factor in lowering the effectiveness of the oxygen sink at the Archaean–Proterozoic transition. Quite simply, the surface sink provided by the reaction of iron (II) to iron (III) might have diminished, as basalt (an Fe(II)-rich rock) became a subordinate part of the material subjected to weathering. The process is independent of

Animal behaviour

Greensleaves

Bats are usually thought of as roosting in caves or buildings. But there are several do-it-yourself tropical species that create tents out of large leaves. They do this by biting the leaf in such a way that the sides collapse downwards and create a sheltered area — as shown here (from within, complete with bat occupants).

These tents can stay healthy for long periods, even though much of the leaf's water supply has been

cut off. How? Carol A. Peterson, Brock Fenton and colleagues have tackled the question by looking at the leaves of three plant species used by bats in Costa Rica (*Biol. J. Linn. Soc.* **72**, 179–191; 2001).

The group examined the water-conduction system in the leaves, and the damage caused to it by bats, and also performed tests with a tracer dye to follow water flow in damaged and undamaged tissue. The vein system in the

leaves is complex, consisting of an apparent hierarchy of some five conducting elements. But it seems that water conduction by the smallest elements is enough to sustain the leaf, even when more major veins are severed. The authors point out that the advantage for the leaf is that it can survive damage from, say, wind or heavy rain; for the bats, it means that they are not forever making new tents.

Tim Lincoln



MICHAEL & PATRICIA FOGDEN/CORBIS

the precise oxidation state of the mantle as all basalts are Fe(II)-rich. It does, however, depend on the volcanic and tectonic processes in the mantle that control the amount of basalt that ascends to the surface where it can be weathered and hence oxidized.

It was over two centuries ago that Antoine Lavoisier figured out that we breathe oxygen. But we still don't know how an oxygen-rich atmosphere arose. Clearly, processes at both the Earth's surface and in its bowels were involved. Exactly how and by how much each contributed remain open questions, but

Kump and colleagues' stimulating ideas should prompt further investigation. ■

Norman H. Sleep is in the Department of Geophysics, Stanford University, Stanford, California 94305, USA.

e-mail: norm@pangea.stanford.edu

1. Kump, L. R., Kastig, J. F. & Barley, M. E. *Geochem. Geophys. Geosyst.* (2001).
<http://www.146.201.254.53/publicationsfinal/researchletters/2000GC000114/fs2000GC000114.html>
2. Langan, R. T. & Sleep, N. H. *J. Geophys. Res.* **87**, 9225–9235 (1982).
3. Garrels, R. M. & Perry, E. A. Jr in *The Sea* Vol. 5 (ed. Goldberg, E. D.) 303–336 (Wiley, New York, 1974).
4. Godderis, Y. & Veizer, J. *Am. J. Sci.* **300**, 434–461 (2000).

Cognitive neuroscience

Repression revisited

Martin A. Conway

The idea that unwanted memories can be repressed has been controversial. An adaptation of an old technique provides an unambiguous model for exploring memory repression in the laboratory.

The concept of repressing unwanted memories is central to psychoanalytic theory, and Freud, in a definition memorable for its clarity, wrote “the essence of repression lies simply in turning something away, and keeping it at a distance from the conscious”¹. Freud provided many examples of memory repression from clinical cases, and documented its effects in day-to-day behaviour². Yet evidence from case studies and vignettes from everyday life are often open to alternative interpretations, and may not be accepted as credible scientific data because they arise from ‘uncontrolled’ conditions. What was needed was a way to investigate memory repression in the laboratory, and Anderson and Green provide just that on page 366 of this issue³. Their methods offer a way of exploring the underlying inhibitory mechanisms, and may ultimately shed light on how repression comes about.

Attempts to study repression in the 1960s and early 1970s produced mixed results, sometimes supporting the idea that unwanted memories can be repressed, and sometimes not. By contrast, more general experimental studies have consistently shown that powerful inhibitory processes are at work in other aspects of human memory⁴. For example, there are inhibitory mechanisms that support the selection of competing items stored in memory.

Anderson and Green's procedure³ is an elegant way of experimentally inducing memory repression. They adapted the well-known ‘go/no-go’ procedure used to study so-called executive control — which regulates the encoding, manipulation and recall of information in working memory — of movements in primates. In the authors' adaptation, human participants first learned a list of pairs of unrelated words (paired associates), such

as ‘ordeal’ and ‘roach’. They then undertook a ‘think/no-think’ task. They were presented with one of the words from a previously learned pair, and asked either to say aloud the associated word (‘think’) or to avoid thinking about it (‘no-think’). So, the no-think condition had two components that had to be inhibited: thinking about the unrepresented word, and saying it aloud. Before the words were presented, the participants were told which words would cue avoidance and which would not, so that they knew whether, on being shown the word ‘ordeal’, for example, they were to say aloud ‘roach’ or to avoid thinking about it. Different paired associates were tested a varying number of times (1, 8 or 16 times, or not at all for ‘baseline’ words).

Some time after the think/no-think phase, participants were presented again with all the cues seen in the previous phase, as well as with cues from baseline pairs. They were then asked to recall the words with which the cues had originally been paired, regardless of whether they had been told to avoid or recall those words in the think/no-think phase. Not surprisingly, participants fared best at recalling words from the ‘think’ category; paired associates that had been practised most often (that is, in 16 trials) were recalled best of all. The next best recall was of the baseline paired associates. Words in the ‘no-think’ category were recalled badly, with the words avoided most often being recalled worst of all. In other words, the more rehearsal, the better the recall (this has been found previously), and the more active the avoidance of rehearsal, the greater the inhibition of recall.

Next, Anderson and Green presented the participants with cues related to the paired associates, and asked them to recall the relevant word. For example, instead of

‘ordeal-r_____’ to cue recall of ‘roach’, the participants were presented with ‘insect-r_____’. Again, words in the no-think category of the original test were poorly recalled, and again, inhibition of recall was strongest for those items for which avoidance had been most practised. This crucial finding shows that it is the avoided word (‘roach’, in this example) that is inhibited, rather than the whole paired associate.

In short, Anderson and Green have shown in a laboratory setting that if a memory that is associated with something familiar (here a word) is actively avoided every time that familiar object is seen, then the memory becomes repressed and the avoided item is later difficult to remember. This is an important result, which lends support to Freud's original definition of repression: it unambiguously shows the existence of consciously initiated, executive inhibition of memory. Even more surprising is that this occurs for unrelated pairs of words — hardly comparable to psychodynamic (primitive) motives or to the amnesia that can result after the trauma of childhood abuse. The results therefore further demonstrate the ubiquity of inhibitory processes in human memory.

More generally, memory researchers⁴ have concluded that inhibition is not only widespread but also habitual, and that, when an item is retrieved from memory, associated items are temporally inhibited. Indeed, the functional arrangement of the brain also suggests the widespread existence of inhibition. As many as 20–25% of cells in the cerebral cortex may be inhibitory and serve to prevent other neurons from becoming overexcited⁵. Such networks of inhibitory neurons could give rise to the type of repression proposed by Freud to underlie neuroses⁶.

In this respect, it is interesting that Anderson was led to his present research by a cognitive analysis of a pattern of amnesia seen in abused children⁷. Children abused by a trusted caregiver are more likely eventually to forget the abuse than those maltreated by strangers. This led to the insight that a trusted caregiver might represent an unavoidable cue for memory retrieval. The only way to prevent persistent recall of damaging memories would be to adapt internally and to deliberately avoid thinking of such memories — in Freud's terms, to push them away from consciousness. Anderson and Green have now shown³ that, even in the innocuous setting of the laboratory, and with stimuli as trivial as randomly paired words, powerful inhibition can be evoked. How much stronger must this inhibition be for objects central to our thoughts and emotions. ■

Martin A. Conway is in the Department of Experimental Psychology, University of Bristol, 8 Woodland Road, Bristol BS8 1TN, UK.
e-mail: m.a.conway@bristol.ac.uk

1. Freud, S. *The Standard Edition of the Complete Psychological Works of Sigmund Freud* Vol. 14 (transl. Baines, C. M. &

- Strachey, J., ed. Strachey, J.) (Hogarth, London, 1957).
2. Freud, S. *The Standard Edition of the Complete Psychological Works of Sigmund Freud* Vol. 6 (transl. Strachey, J.; ed. Strachey, J.) (Hogarth, London, 1960).
3. Anderson, M. C. & Green, C. *Nature* **410**, 366–369 (2001).
4. Bjork, R. A., Bjork, E. L. & Anderson, M. C. in *Intentional Forgetting: Interdisciplinary Approaches* (eds Golding, J. M. & MacLeod, C. M.) 103–137 (Lawrence Erlbaum, Mahwah, NJ, 1998).
5. Houghton, G. & Tipper, S. T. *Brain Cognition* **30**, 20–42 (1996).
6. Freud, S. *The Standard Edition of the Complete Psychological Works of Sigmund Freud* Vol. 1 (transl. Strachey, J.; ed. Strachey, J.) (Hogarth, London, 1957).
7. Anderson, M. C. *J. Aggression Maltreatment Trauma* (in the press).

Superconductivity

Magnetic glue exposed

Piers Coleman

Overcoming electrons' mutual repulsion is the key to superconductivity. In simple metals, the forces that bind these charged particles in pairs are well understood, but what is the glue in other superconductors?

Physicists are fascinated by the 'glue' that holds matter together. According to quantum mechanics, forces are not instantaneous, but are transmitted by the exchange of tiny packets of energy, or 'quanta', between particles. So electromagnetic forces are produced by the exchange of photons, and the strong forces that keep quarks tightly bound inside protons are driven by the exchange of quanta suggestively called gluons.

But inside metals, different sorts of quanta become possible, and electrons that exchange these quanta can experience new kinds of attractive forces that profoundly change their properties. The most celebrated example of this is superconductivity — which occurs when electrons at low temperature pair up and flow without resistance. Now, on page 340 of this issue, Sato *et al.*¹ reveal a magnetic origin for the 'glue'

between electrons in an unconventional superconductor.

In the late 1950s, Bardeen, Cooper and Schrieffer (BCS) showed that superconductivity involves the formation of bound pairs of electrons, named Cooper pairs. BCS argued that the electron pairs were 'glued together' by tiny deformations in the crystal lattice, called phonons, that accompany the electrons' motion. But although phonons were implicated in superconductivity many years before the BCS theory, it was not until the 1960s that it became possible to definitively identify them as the glue in conventional superconductivity.

In the early 1960s, the Russian physicist Eliashberg² showed how the electron-pairing forces created by phonons could be elegantly incorporated into BCS theory using a set of equations that now bear his name. It turns out that the exchange of phonons between

the electrons in a Cooper pair produces tiny harmonics — wiggles and bumps — in the energy distribution of the electrons that can be detected using electron-tunnelling spectroscopy³. In the late 1960s, McMillan and Rowell⁴ confirmed the long suspected role of phonons in conventional superconductors. They showed that the phonon energy spectrum measured by neutron scattering from superconducting mercury and lead agreed with the spectrum that they had deduced from their precise electron-tunnelling measurements. In so doing, they also confirmed that the Eliashberg refinement of BCS theory was accurate to about 1%.

The 1970s and 1980s led to the discovery of superconductivity in two new, unconventional classes of material, the heavy-fermion^{5,6} and high-temperature⁷ superconductors. Conventional superconductivity is suppressed by the tiniest concentration of magnetic atoms, but the unconventional superconductors contain a dense array of magnetic atoms, which appear to be actively involved in electron pairing. This led physicists to consider a new kind of 'magnetically mediated' superconductivity in which the quanta that glue electrons into pairs are derived from magnetic fluctuations^{8–12}.

Like electrons in atoms, the character of the electron waves in Cooper pairs is labelled *s*, *p* or *d*, depending on the angular momentum of the electrons. Phonon-mediated pairing favours the formation of *s*-wave electron pairs, but magnetic fluctuations favour *p*- or *d*-wave pairs (Fig. 1). This is because magnetic fluctuations develop in metals with strongly interacting electrons: *p*- or *d*-wave pairs form in direct response, lowering the mutual repulsion between the electrons by keeping them further apart. In the unconventional superconductors, neutron scattering experiments have revealed large fluctuations in magnetic spin at low temperatures, which could mediate electron pairing, and several other experiments provide good evidence for the existence of *p*- or *d*-wave electron pairs in these materials.

But despite strong circumstantial evidence for magnetically mediated superconductivity, definitive proof that spin fluctuations are the glue for *p*- or *d*-wave pairing has been hard to come by. This is partly because magnetic fluctuations are far more difficult to characterize than phonons, but also because, unlike phonons, magnetic fluctuations interact with each other, making the theoretical description of magnetically mediated pairing a far more complex problem.

Sato *et al.*¹ provide new evidence for this magnetic glue. They studied the heavy-fermion superconductor UPd₂Al₃, so called because the conducting electrons in it acquire very large effective masses, often hundreds of times that of a free electron. UPd₂Al₃ is one of a handful of heavy-fermion superconductors in which both antiferro-

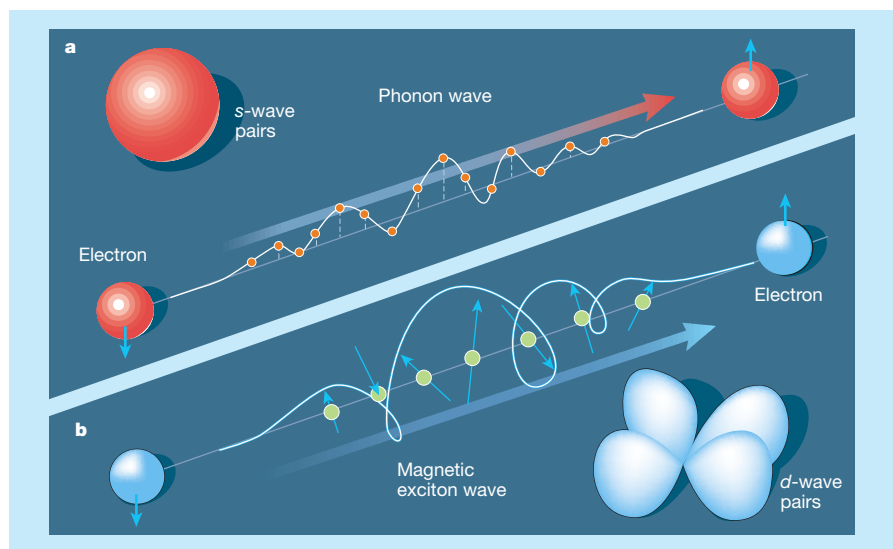


Figure 1 Two ways to make a metallic superconductor. a, In conventional superconductivity the glue responsible for binding electrons into superconducting pairs is derived from the exchange of 'quanta' of lattice vibrations — phonons, shown here as a wave passing through the lattice atoms — that bind electrons into 's-wave' pairs. b, Sato *et al.*¹ provide new evidence that, in the unconventional superconductor UPd₂Al₃, the glue responsible for the electron pairing derives from the exchange of magnetic quanta — magnetic excitons — that bind the electrons into 'd-wave' pairs (blue arrows indicate direction of spin).

magnetism (a type of magnetic order) and superconductivity can coexist. In essence, the work by Sato *et al.* is an attempt to extend the approach of McMillan and Rowell to unconventional superconductivity.

In earlier neutron scattering experiments on UPd₂Al₃, Sato and colleagues¹³ detected an excitation in the material's magnetic spin fluctuations, which they identified as a 'magnetic exciton'. Magnetic excitons are a kind of spin wave that ripples through a magnetically ordered medium, but they require a certain amount of energy to become excited. Sato *et al.*¹³ noticed that in the normal metallic state, the magnetic exciton is strongly scattered by the heavy electrons in UPd₂Al₃ and has a short lifetime, but that in the superconducting state, it is weakly scattered, with a long lifetime. This suggested that excitation is strongly coupled to the electrons involved in pairing, and it led Sato and colleagues to speculate that the magnetic exciton might indeed be the long-sought glue particle.

To test this idea, Sato *et al.*¹ built a simple model in which the magnetic exciton couples magnetically to the heavy electrons. They found that, with strong coupling, they could account for the dramatic reduction in the excitons' lifetime in the normal metallic state. Their model also explained the appearance of a secondary peak in the exciton spectrum in the superconducting state as the result of an energy gap developing in the heavy-electron spectrum. Finally, in a step that closely parallels the work of McMillan and Rowell, Sato *et al.* used the Eliashberg equations to predict the electron energy spectrum expected from neutron tunnelling measurements. Their model can explain both the observed superconducting transition at a temperature of 2 K, and the fine details of the tunnelling spectra in the superconducting state.

These fascinating results argue strongly that magnetic fluctuations are indeed the glue that drives *d*- or *p*-wave pairing in this unconventional superconductor. But there are still many open questions. UPd₂Al₃ is a rather special case of *d*- or *p*-wave pairing because the electrons move in a magnetically ordered medium. It is this feature that allows the spin fluctuations to move ballistically, enabling them to be regarded like phonons or their free-space cousin, the photon.

The situation in many other unconventional superconductors is not as clear-cut. For instance, in the high-temperature superconductors and many heavy-fermion superconductors, there is no fully developed antiferromagnetic order, and there is no clear separation between the magnetic spin fluctuations and conducting electrons. Here, the glue is much more likely to resemble the 'gluons' that bind quarks inside nucleons, in that the interactions are far stronger and the glue particle only exists as a fleeting entity, rather than as a well-defined excitation. This is the

picture that is emerging to explain similar magnetic excitons that have been seen in high-temperature superconductors¹⁴. Nonetheless, the evidence provided by Sato *et al.* marks an important step forward: introducing us to a material in which magnetically mediated pairing plays a similar role to conventional phonon-mediated pairing. ■

Piers Coleman is in the Department of Physics and Astronomy, Rutgers University, 136 Frelinghuysen Road, Piscataway, New Jersey 08854-8019, USA. e-mail: coleman@physics.rutgers.edu

1. Sato, N. K. *et al.* *Nature* **410**, 340–343 (2001).
2. Eliashberg, G. *Zh. Eksp. Teor. Fiz.* **38**, 966–976 (1960); transl. in *Sov Phys JETP* **11**, 696–702 (1960).

Structural biology

Protein crystal mimics reality

Robert M. Macnab

Flagellin, a protein in the bacterial flagellar motor, has resisted attempts at three-dimensional crystallization — until now. The crystal structure offers hints as to how filaments are able to switch helical states.

Bacteria such as *Salmonella* and *Escherichia coli* propel themselves through aqueous environments using flagella. The flagellum has two parts: the filament and the motor. The filament is a long, thin, helical structure found outside the cell; when rotated by the motor, embedded in the cell surface, the filament propels the cell along. The filament is made from tens of thousands of copies (subunits) of one protein, flagellin. Changes in the arrangement of these subunits within the filament result in changes in bacterial movement, from 'swimming' to 'tumbling', but how does this molecular switch come about? On page 331 of this issue¹, Samatey and colleagues provide the most detailed picture to date of this interesting biological structure, and make a prediction about the switching event.

Flagellar filaments are organized as follows. Flagellin proteins are arranged, one on top of another, into a protofilament, and 11 of these protofilaments are lined up side by side, forming what looks like a loosely rolled sheet of paper, or hollow cylinder — the flagellar filament (Fig. 1a, overleaf). The protofilaments can be in either of two states, called L (left-handed) and R (right-handed), with all subunits within a given protofilament being in the same left- or right-handed state. The distance between a flagellin subunit and that above or below it (the intersubunit distance) is almost, but not exactly, the same in the L and R states (52.7 Å and 51.9 Å, respectively).

In the abnormal situation in which all 11 protofilaments in a filament are in the same state, the filament is straight. But usually it contains protofilaments in both states, and so is macroscopically helical — an essential feature for propulsion. For example, in the case

3. Rowell, J. M., Anderson, P. W. & Thomas, D. E. *Phys. Rev. Lett.* **10**, 334–335 (1963).
4. McMillan, W. L. & Rowell, J. M. in *Superconductivity* Vol 1 (ed. Parks, R. D.) 561–611 (Dekker, New York, 1969).
5. Steglich, F. *et al.* *Phys. Rev. Lett.* **43**, 1892–1895 (1976).
6. Ott, H. R., Rudigier, H., Fisk, Z. & Smith, J. L. *Phys. Rev. Lett.* **50**, 1595–1598 (1983).
7. Bednorz, J. G. & Müller, K. A. *Z. Phys. B* **64**, 189–193 (1986).
8. Anderson, P. W., Brueckner, K. A., Soda, T. & Morel, P. *Phys. Rev.* **118**, 1442–1446 (1960).
9. Emery, V. & Sessler, A. M. *Phys. Rev.* **119**, 43–49 (1960).
10. Miyake, K., Schmitt Rink, S. & Varma, C. M. *Phys. Rev. B* **34**, 6554–6556 (1986).
11. Béal-Monod, M. T., Bourbonnais, C. & Emery, V. *Phys. Rev. B* **34**, 7716–7720 (1986).
12. Bedell, K. S. & Pines, D. *Phys. Rev. B* **37**, 3730–3733 (1988).
13. Sato, N. K. *et al.* *J. Phys. Soc. Jpn* **66**, 1884–1887 (1997).
14. Zasadzinski, J. F. *et al.* <http://xxx.lanl.gov/abs/cond-mat/0102475>

of the form used for swimming², there are nine L and two R protofilaments. The protofilaments have a remarkable capacity to switch between different states in a cooperative fashion. This manifests itself at the macroscopic level as conversion between different helical forms, notably the left-handed form used for swimming (when the motor rotates in an anticlockwise direction) and the right-handed form used for reorientation (tumbling, when the motor rotates clockwise)³.

High-resolution structural analysis of a three-dimensional crystal of flagellin would help to explain how this molecular switch takes place. But the subunits of filamentous structures often have a mind of their own — they prefer to organize themselves in a long line, the filament's axis. Flagellin is a perfect example. *In vitro*, flagellin can readily form filaments, identical to those formed *in vivo*. But, until now, no one had succeeded in obtaining a three-dimensional crystal of flagellin, and so structural studies^{4,5} — even using the most advanced techniques of electron microscopic image analysis and X-ray-fibre diffraction — gave a resolution of only about 10 Å. This is good enough to yield information about the shape and overall organization of the subunits. But it does not give the atomic resolution (2 Å or better) needed fully to understand a protein's secondary structure (how particular regions fold, for example into α -helices) and tertiary structure (overall conformation). Nor can it provide detailed information about quaternary structure — the interfaces between proteins — which is so important in a structure such as the flagellar filament.

Samatey *et al.*¹ describe how they precisely and judiciously cut off the amino- and carboxy-terminal ends of the flagellin, so that

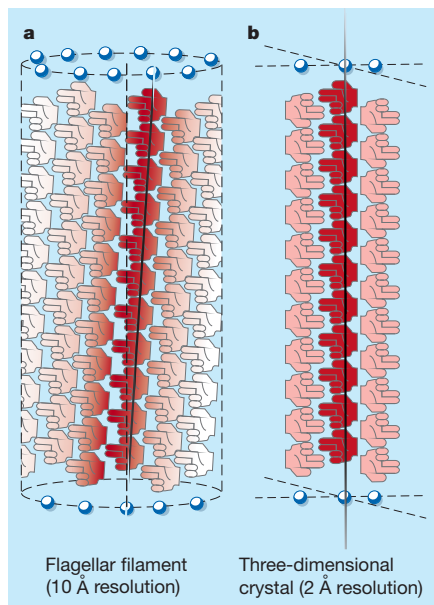


Figure 1 A comparison of the way in which flagellin proteins (represented by hands) incorporate themselves into the flagellar filament that propels bacteria, and into a three-dimensional crystal. a, The flagellar filament, understood at the relatively low resolution of about 10 Å, is a cylindrical structure, which can be thought of as a set of 11 protofilaments (centred on the blue circles) arranged in parallel and slightly tilted with respect to the filament's axis (dashed line). One protofilament is emphasized by a darker shade and a line indicating its direction. b, To obtain a three-dimensional crystal, Samatey *et al.*¹ used a version of flagellin that was truncated at both ends to inhibit filament formation. The subunits on one of the three crystal axes correspond closely to those in protofilaments, although now arranged in antiparallel fashion. The authors built this high-resolution (2 Å) structure into the lower-resolution filament structure. Only slight adjustments had to be made for the fact that the protofilaments in the filament are helices, gradually winding around the particle axis, whereas the protofilaments in the crystal are straight.

the protein's ability to form filaments was removed, yet its structure was largely retained. The authors then persuaded the truncated protein to form three-dimensional crystals. The crystals were exceedingly thin, being only a few micrometres wide. This presented technical difficulties that could be overcome only by using some of the most powerful beam lines available to generate X-rays for analysis, and by taking advantage of recent advances in low-temperature crystallography and data analysis. Samatey *et al.* thereby succeeded in obtaining data at atomic resolution (2.0 Å).

An interesting result emerged. The intersubunit distance on one of the three crystal axes was 51.8 ± 0.1 Å — more or less identical to the intersubunit distance of 51.9 Å within an R-state protofilament. This was unlikely to be a coincidence, and with only minor adjust-

ments, images of this high-resolution structure (Fig. 1b) could be built into images of the lower-resolution filament structure obtained by electron microscopy. In the natural filament, the protofilaments are parallel, whereas in the crystal they are antiparallel. So the lateral interactions between protofilaments in the filament are entirely different in the two cases. Yet the fact that the protofilaments look so similar means that lateral interactions, though undoubtedly important for stability, are not essential for the fundamental organization of a protofilament, which is therefore likely to be the cooperative unit for switching.

The flagellin that Samatey *et al.* crystallized had the information at its amino- and carboxy-termini deleted. Could this be resulting in a false picture? This seems unlikely because of how well the crystal structure fits into the lower-resolution structure of the filament, which is made from full-length flagellin. Also, other studies indicate that the termini lie towards the inside of the hollow filament, contributing to its strength and rigidity, but probably not to its switching capability.

For technical reasons, analysis of flagellar filaments is always restricted to mutant forms in which all 11 protofilaments are in the same state, either L or R. Oddly, Samatey *et al.* found that both L and R flagellins (that is, flagellins that form L and R protofilaments, respectively) crystallized into the R state. This would have been a disappointment if the authors wished (as surely they must have) to look directly at the molecular switch between states. In the hope of simulating the slightly longer L state, they used a computer to gradually stretch their R-type protofilament images, and slightly adjusted the structure obtained at each step so as to minimize its energy. For several steps, nothing happened bar (for the most part) the gradual stretching of α -helices. But at one point, a local feature of the structure, at the intersubunit interface, changed abruptly, slightly distorting the protofilament. The authors suggest, quite plausibly, that this may be the key event in switching.

Samatey *et al.*'s work¹ represents a major jump forwards in our understanding of this sophisticated structure and the protein subunit from which it is built. Yet, for a full understanding of how the filament switches between states, one really needs high-resolution crystal structures of both states. Crystallization conditions that make this possible may eventually be found.

Robert M. Macnab is in the Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, PO Box 208114, New Haven, Connecticut 06520-8114, USA. e-mail: robert.macnab@yale.edu

1. Samatey, F. A. *et al.* *Nature* **410**, 331–337 (2001).
2. Calladine, C. R. *J. Mol. Biol.* **118**, 457–479 (1978).
3. Macnab, R. M. & Ornston, M. K. *J. Mol. Biol.* **112**, 1–30 (1977).
4. Mimori, Y. *et al.* *J. Mol. Biol.* **249**, 69–87 (1995).
5. Morgan, D. G. *et al.* *J. Mol. Biol.* **249**, 88–110 (1995).

Daedalus

A near miss

In the last days of the year 2000, a meteor (later designated Y2000) passed within 480,000 miles of Earth, and so had a chance of a few parts per million of a direct hit. Daedalus is more interested in an event of 10 August 1972, when a meteor weighing maybe 4,000 tons skimmed through Earth's atmosphere, coming to within 58 kilometres over the United States and Canada. Daedalus reckons that such a close entry into the resistive atmosphere must have slowed the object to less than orbital velocity, so that its subsequent trajectory will have intercepted the Earth, possibly over land, but more likely in the sea. If it had hit the land it would surely have been noticed.

Two-thirds of meteors hit the sea anyway; and indeed sea impacts have been blamed for many important events, such as the extinction of the dinosaurs. Yet no sea impact has ever been observed and its remains studied. The chance should surely be taken. So Daedalus recommends that the trajectory of the 10 August 1972 meteor should be carefully analysed. The results will determine where in the ocean the incomer splashed down — somewhere on the great circle extension of the trajectory seen over North America. The area of greatest probability will be an ellipse, very likely over some area of ocean.

DREADCO oceanographers will then study that area, using submersible craft equipped with very sensitive cameras. Daedalus regrets the use of powerful lights by such craft, as they blind all the poor sea-creatures that have evolved sensitive retinas. The workers could also look for an altered magnetic signature of the ocean floor. Daedalus has no idea what his oceanographers will make of the results. In principle, a large hot object hitting the sea should make a giant boiling splash, killing all sea life within a large radius. The remains should then descend to the seabed, which might be deep ooze or teeming continental shelf. Because the impact was only 29 years ago, the main remnant should still be interpretable, even if it is only a magnetic anomaly.

It is possible, of course, that the bolide exceeded impact velocity even after its encounter with the atmosphere — in which case it may still be in low Earth orbit, waiting for a chance to deliver its deadly warning again. But then Daedalus would expect that Earthly radio detection or radar, which failed so signally to detect the near miss of 2000, would have picked it up by now, and would predict a good impact site.

David Jones

Frequency-dependent Batesian mimicry

Predators avoid look-alikes of venomous snakes only when the real thing is around.

Batesian mimicry holds that palatable species look like dangerous species because both are then protected from predation^{1–5}. But this protection should break down where the dangerous model is absent, when predators would not be under selection to recognize the model or any other species resembling it as dangerous^{2,4,5}. Here we provide experimental evidence to support this critical prediction of Batesian mimicry by demonstrating that predators avoid harmless look-alikes of venomous coral snakes only in areas that are inhabited by these deadly snakes.

Many coral snakes and non-venomous kingsnakes possess red, yellow (or white), and black ringed markings⁶, which predators avoid⁷, though often without prior experience⁸. To determine whether this avoidance depends on the model's presence in the vicinity, we constructed snake replicas⁷ (1.5 cm × 18 cm cylinders of pre-coloured, non-toxic plasticine threaded onto an S-shaped wire) with a tricolour ringed pattern, a striped pattern with identical colours and proportions as the ringed replicas, or a plain brown pattern.

Ringed replicas conformed to the local mimic: scarlet kingsnakes (*Lampropeltis triangulum elapsoides*), which resemble eastern coral snakes (*Micrurus fulvius*)⁹, or sonoran mountain kingsnakes (*L. pyromelana*), which resemble western coral snakes (*Micruroides euryxanthus*)¹⁰; striped and brown replicas served as controls. We arranged three different replicas (triplets) 2 m apart in natural habitat (each was used once only). At each site, 10 triplets were placed 75 m apart in a line. After collection, a person without knowledge of the replica's location scored attacks by noting any impressions corresponding to a predator⁷.

We tested whether predators avoid *L. t. elapsoides* only in areas inhabited by *Micrurus* by placing 10 triplets at eight sympatric sites (sites with mimic and model) and eight allopatric sites (sites with only the mimic) in North and South Carolina, USA (480 replicas; allopatric sites were more than 80 km outside *Micrurus*'s range^{9,11}; sites were 16–420 km apart). After 4 weeks, 25 (5.2%) replicas had been attacked by carnivores. The mean ($\pm$ s.e.m.) proportion of ringed replicas attacked was significantly greater in allopatry (0.654 ± 0.107) than in sympatry (0.083 ± 0.116 ; $P = 0.009$, 2-tailed Wilcoxon two-group test).

We next investigated whether predators avoid *L. pyromelana* only in sympatry with *Micruroides* by placing 10 triplets at 24 sites (720 replicas) along an elevational gradient

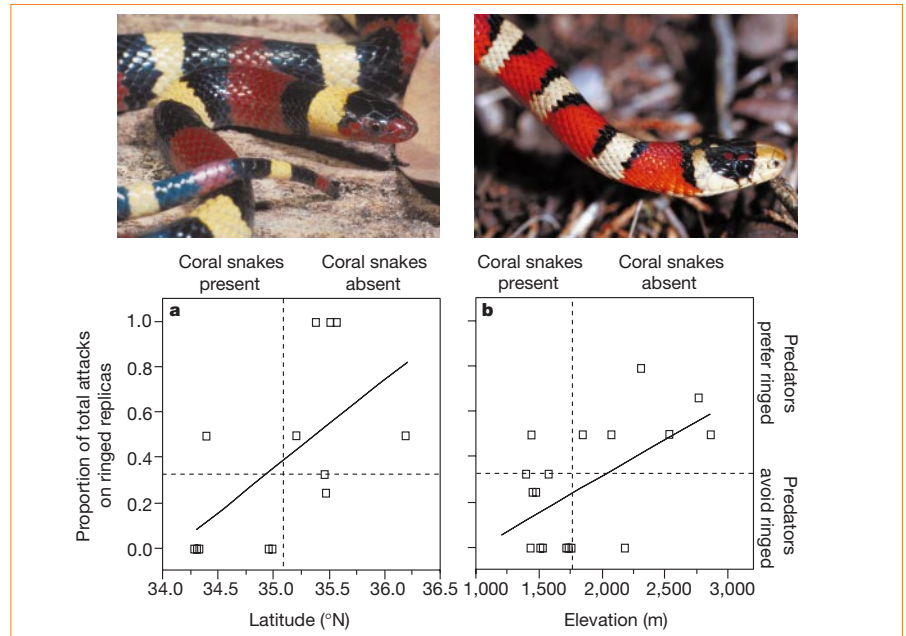


Figure 1 Frequency-dependent mimicry. The proportion of carnivore attacks on ringed replicas of scarlet kingsnakes (top left; a mimic of eastern coral snakes) and sonoran mountain kingsnakes (top right; a mimic of western coral snakes) increased with **a**, latitude ($y = -13.314 + 0.391x$, $P < 0.035$, $R^2 = 0.345$) and **b**, elevation ($y = -0.329 + 0.00032x$, $P < 0.014$, $R^2 = 0.310$). Horizontal dashed line: proportion of attacks on ringed replicas expected under randomness. Vertical dashed line: maximum latitude and elevation for coral snakes in North Carolina and Arizona, respectively.

(1,204–2,866 m) near Portal, Arizona (*Micruroides* only occur at altitudes below 1,770 m (ref. 10); there were 14 sympatric and 10 allopatric sites 3–100 km apart). After 2 weeks, 49 (6.8%) replicas had been attacked by carnivores.

The mean proportion of ringed replicas attacked was significantly greater in allopatry (0.496 ± 0.078) than in sympatry (0.138 ± 0.060 ; $P = 0.006$). Moreover, in sympatry, the proportion of ringed replicas attacked (0.138) was significantly less than randomness (0.33; $P = 0.010$, 2-tailed Wilcoxon signed-rank test). By contrast, attacks were random in allopatry ($P = 0.188$). Thus, predators avoid coral snake mimics only in sympatry with the model.

Coral snakes become increasingly rare with increasing latitude (Spearman $\rho = -0.57$, $P = 0.014$)¹¹ and elevation ($\rho = -0.77$, $P = 0.026$; our unpublished results). Consequently, selection to avoid ringed patterns should weaken with increasing latitude and elevation. As expected, the proportion of ringed replicas attacked increased gradually with latitude and elevation (Fig. 1), suggesting that selection to avoid ringed patterns is indeed sensitive to the abundance of coral snakes.

Our results do not fully resolve why mimetic patterns occur where models are absent^{6,9–11}. Possibly selection for mimicry

in sympatry, coupled with gene flow between sympatric and allopatric populations¹², maintains mimetic patterns in both regions. Nevertheless, our results verify the critical prediction of Batesian mimicry and demonstrate that the benefits of mimicry depend on abundance of the model.

David W. Pfennig[†], William R. Harcombe^{*}, Karin S. Pfennig[†]

^{*}Department of Biology, University of North Carolina, Chapel Hill,

North Carolina 27599-3280, USA

e-mail: dpfennig@email.unc.edu

[†]Section of Integrative Biology, University of Texas, Austin, Texas 78712-1064, USA

1. Bates, H. W. *Trans. Linn. Soc. Lond.* **23**, 495–566 (1862).
2. Wallace, A. R. *Contributions to the Theory of Natural Selection* (Macmillan, London, 1870).
3. Fisher, R. A. *The Genetical Theory of Natural Selection. A Complete Variorum Edition* (Oxford University Press, New York, 1999).
4. Waldbauer, G. P. *Evol. Biol.* **22**, 227–259 (1988).
5. Maynard Smith, J. *Evolutionary Genetics* (Oxford University Press, Oxford, 1998).
6. Greene, H. W. & McDiarmid, R. Y. *Science* **213**, 1207–1212 (1981).
7. Brodie, E. D. *Evolution* **47**, 227–235 (1993).
8. Smith, S. M. *Science* **187**, 759–760 (1975).
9. Conant, R. & Collins, J. T. *A Field Guide to Reptiles and Amphibians of Eastern and Central North America* (Houghton Mifflin, Boston, MA, 1998).
10. Stebbins, R. C. *Western Reptiles and Amphibians* (Houghton Mifflin, Boston, MA, 1985).
11. Palmer, W. M. & Braswell, A. L. *Reptiles of North Carolina* (University of North Carolina Press, Chapel Hill, NC, 1995).
12. King, R. B. & Lawson, R. *BioScience* **47**, 279–286 (1997).

Pattern formation

Instabilities in sand ripples

Sand ripples are seen below shallow wavy water and are formed whenever water oscillates over a bed of sand. Here we analyse the instabilities that can upset this perfect patterning when the ripples are subjected to large changes in driving amplitude or frequency, causing them to deform both parallel and transverse to their crests. Our results reveal new pattern-forming instabilities in granular matter exposed to fluid flow with strong vorticity.

The formation of normal ripple patterns is driven by the 'separation zones' (regions with reversed flow) created in the troughs between ripple crests^{1,2}. The wavelength of the ripples is set by the size of these zones, which is roughly proportional to the amplitude of the fluid oscillations and is independent of the frequency^{3–5}. For a typical driving of amplitude 2.8 cm and frequency 0.5 Hz, for example, a ripple pattern with wavelength close to 4.0 cm is formed.

To investigate ripple patterns, we placed a thick layer of sand (spherical glass beads of diameter 250–350 μm) on a 0.6 m $\times$ 1 m tray inside a large, closed, stationary water tank with a transparent top lid, and pulled the tray back and forth sinusoidally. We started with completely regular patterns formed by pressing down a mould in the flat bed and then running the system (for a

short time) with an amplitude 'commensurate' with the imprinted ripple wavelength.

We 'frustrated' these perfect patterns by suddenly changing the amplitude or frequency of the oscillations. When these changes are not too large, periodic ripple patterns are stable. For example, for a frequency of 0.67 Hz, ripples of wavelength 4.2 cm are stable for driving amplitudes between 2.4 and 3.4 cm. But when the driving amplitude or frequency is quenched beyond some threshold, three distinct short-wavelength instabilities occur.

First, if the driving amplitude is increased sufficiently, the ripple crests distort along and transverse to the original ripples, and the pattern of compressed and 'bulging' regions forms a checkerboard tilted by 45° (Fig. 1a). As the deformations grow, dislocations occur until finally a new periodic ripple pattern with a larger wavelength is formed. Second, if the transition is reversed by decreasing the amplitude sufficiently, the pattern 'doubles' through the generation of small ripples between each of the original ones (Fig. 1b). Here, the separation zones become so small that they do not reach the troughs between the ripples. After the doubling, the wavelength of the system is then too small and the final state is reached through a bulging instability. Third, if the driving frequency is increased sufficiently, the ripples become unstable to the formation of stationary 'pearls' in the troughs between the ripples, which again form a tilted pattern (Fig. 1c). Unlike bulging and

doubling, pearling saturates and disappears reversibly when the frequency is decreased.

Although our system superficially resembles other classic pattern-forming systems⁶ such as Rayleigh–Bénard convection, the transition scenarios described here have not been seen before. We believe that the theory of these instabilities must start with the separation zones, which have the shape of cylindrical vortices along the ripple troughs. These might be prone to instabilities⁷ like the Rayleigh–Plateau 'sausage' instability (which may lead to bulging), or the centrifugal instability that gives rise to Taylor vortices whose axes are transverse to the cylinder (which may lead to pearling).

Jonas Lundbek Hansen*, Martin van Hecke*, Anders Haaning*, Clive Ellegaard*, Ken Haste Andersen†, Tomas Bohr‡, Thomas Sams†

*Niels Bohr Institute, Blegdamsvej 17, 2100 Copenhagen, Denmark

†Danish Defense Research Establishment, Ryvangs allé 1, 2100 Copenhagen, Denmark

‡ISVA, Danish Technical University, 2800 Kgs. Lyngby, Denmark

§Physics Institute, Danish Technical University, 2800 Kgs. Lyngby, Denmark

e-mail: tbohr@nbi.dk

1. Bagnold, R. A. *Proc. R. Soc. A* **187**, 1–15 (1946).
2. Bagnold, R. A. *The Physics of Blown Sand and Desert Dunes* (Chapman & Hall, Methuen, London, 1941).
3. Andersen, K. H. *Ripples beneath Surface Waves*. Thesis, Niels Bohr Inst., Copenhagen (1999).
4. Scherer, M. A. et al. *Phys. Fluids* **11**, 58–67 (1999).
5. Stegner, A. & Wesfried, J. E. *Phys. Rev. E* **60**, 3487–3490 (1999).
6. Cross, M. C. & Hohenberg, P. C. *Rev. Mod. Phys.* **65**, 851–1112 (1993).
7. Drazin, P. & Reid, W. *Hydrodynamic Stability* (Cambridge Univ. Press, Cambridge, 1981).

Palaeoanthropology

Did our ancestors knuckle-walk?

All African apes walk on their knuckles. There is no evidence for this behaviour in the earliest hominids, however, which conflicts with molecular DNA evidence suggesting that chimpanzees are more closely related to humans than to gorillas. On the basis of a multivariate analysis of four traits of the proximal wrist joint, Richmond and Strait¹ claim that African apes and early hominids do share a common knuckle-walking ancestor. I propose that these traits are not uniquely associated with knuckle-walking and question the basis of their conclusion. It is still possible that no human ancestor knuckle-walked and that this behaviour evolved independently in different species.

Although such an ancestor would counter objections to an exclusive human–chimpanzee clade, it would not prove that knuckle-walking evolved only once in the ancestry of African apes and

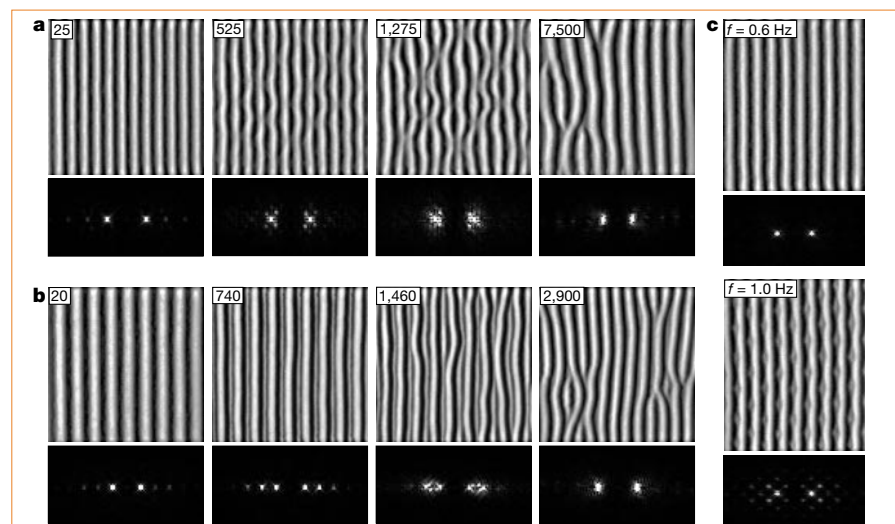


Figure 1 Patterns obtained by changing the driving parameters for a tray of sand oscillating in a water tank. **a, b**, Sequences of pictures from two experiments in which the amplitude of the drive was changed. The tray area in each is 53 cm $\times$ 53 cm; the light source was on the left and the sand tray was oscillated in the left–right direction; the camera was directly above the sand bed. Below each picture is shown the central region of the corresponding power spectrum, $2.32\text{ m}^{-1} \times 1.16\text{ m}^{-1}$. Numbers in the corners show the period at which the pictures were taken. For the top row (**a**), the initial wavelength was 4.2 cm, then the amplitude was 4.0 cm and the frequency 0.41 Hz. The power spectrum shows clear evidence of the bulging modulations mentioned in the text. For the bottom row (**b**), the imprinted wavelength was 5.6 cm, the driving amplitude was 2.8 cm and the frequency 0.68 Hz. The initial doubling is evident. **c**, The pearling transition and the corresponding power spectra in the range $1.88\text{ m}^{-1} \times 0.94\text{ m}^{-1}$. Each picture shows an area of 43 cm $\times$ 53 cm. The ripples have a wavelength of 4.2 cm; the driving amplitude was 2.8 cm; frequencies (f) are indicated. Top, state well below the transition; bottom, fully developed pearling state. The pearls are new small ripples and not distortions of old ripples, because only the troughs and not the crests are deformed. Both states are stable at their respective driving parameters, so no period number is shown.

humans. Richmond and Strait¹ discount the possibility that knuckle-walking evolved independently in gorillas and in a common human–chimpanzee ancestor, an idea supported by the different ontogenetic development of gorilla and chimpanzee carpus load-bearing morphology² and by terrestrial orang-utans who normally walk on the dorsum of their fingers, albeit in a less efficient³ fist-walking posture^{4,5}, and occasionally knuckle-walk^{6,7}. The entire great-ape lineage may thus be predisposed towards knuckle-walking when terrestrial quadrupedalism becomes a major selective factor^{6,7}, undermining the claim that the four cited traits can resolve the phylogenetic relationships of chimpanzees, humans and gorillas¹.

Richmond and Strait¹ assume that because some features of the present-day distal radius fulfil one of the mechanical requisites of African-ape knuckle-walking (limiting extension at the radiocarpal joint), then they must have arisen in response to the requirements of knuckle-walking. But it is debatable whether form, function and adaptive significance should be so linked within a modern evolutionary framework^{8,9}. Claims for knuckle-walking adaptations in the wrist need careful evaluation as this behaviour describes a novel posture of the metacarpals and phalanges in relation to a substrate^{4,5}. Digitigrade, semidigitigrade or fist-walking postures may make similar demands on the wrist joints to knuckle-walking⁴ and so could also have given rise to stabilizing and extension-limiting mechanisms at the radiocarpal joint.

Extension-limiting structures allow contact between the hand and substrate to be maintained more efficiently when climbing large vertical supports⁴ and so may have arisen in a climbing ancestor of humans and great apes¹⁰. From this perspective, Richmond and Strait¹ only demonstrate that behaviours requiring limited extension at the radiocarpal joint were important in hominid ancestry. Methodological considerations, however, suggest that the proposed extension-limiting mechanism of the distal radius may not have been suitably quantified.

A clearly delimited scaphoid notch is not always evident in the hominoid species included in the analysis¹, so the measurements may not be replicable. It is unclear why the angle between the scaphoid and lunate facets of the radius may be significant in quantifying the extension-limiting mechanisms at this joint. This dimension should discriminate between African and Asian apes, with a narrower angle in the latter. The greater radio-ulnar curvature of the distal radial articulation in orang-utans and gibbons is atypical for anthropoid primates, and probably facilitates the rotary movement of the proximal carpal joint needed for forearm suspensory locomotion^{4,11}. But it is not explained why increased rotary

movement should preclude an extension-limiting mechanism at the radiocarpal joint¹. Part of the discrimination between African apes, Asian apes and hominids (their Fig. 2) may therefore reflect differences in suspensory behaviour and have nothing to do with knuckle-walking.

The first two canonical axes of Fig. 2b of ref. 1 do not separate the *A. afarensis* radii from orang-utans any more than from chimpanzees. The variables that strongly contribute to the remaining axis are not identified, so it is unclear why the Mahalanobis D^2 distance is greater between orang-utans and *A. afarensis* than between chimpanzees and *A. afarensis* (their Fig. 2c, based on all canonical axes), a difference that bears on the conclusion that *A. afarensis* had a knuckle-walking ancestry.

Although hominids may have evolved from a knuckle-walker, I do not believe that Richmond and Strait's analysis¹ has proved this. I cannot, therefore, support related interpretations regarding the phylogenetic affinity¹ and locomotor mode¹² of early hominids, or of the chimpanzee-like characteristics of their immediate ancestor¹.

Mike Dainton

Department of Human Anatomy and Cell Biology,
New Medical School, University of Liverpool,
Ashton Street, Liverpool L69 3GE, UK
e-mail: mdainton@liverpool.ac.uk

1. Richmond, B. G. & Strait, D. S. *Nature* **404**, 382–385 (2000).
2. Dainton, M. & Macho, G. A. *J. Hum. Evol.* **36**, 171–194 (1999).
3. Tuttle, R. H. *J. Morphol.* **128**, 309–364 (1969).
4. Sarmiento, E. E. *Int. J. Primatol.* **9**, 281–345 (1988).
5. Tuttle, R. H. *Am. J. Phys. Anthropol.* **26**, 171–206 (1967).
6. Tuttle, R. H. (ed.) in *Primate Functional Morphology and Evolution* 203–212 (Mouton, The Hague, 1975).
7. Tuttle, R. H. & Beck, B. B. *Nature* **236**, 33–44 (1972).
8. Gould, S. J. & Vrba, E. S. *Paleobiology* **8**, 4–15 (1982).
9. Lauder, G. V. in *Functional Morphology in Vertebrate Paleontology* (ed. Thomason, J. J.) 1–18 (Cambridge Univ. Press, Cambridge, 1995).
10. Sarmiento, E. E. *Hum. Evol.* **10**, 289–321 (1995).
11. Jenkins, F. A., Jr & Fleagle, J. G. in *Primate Functional Morphology and Evolution* (ed. Tuttle, R. H.) 213–221 (Mouton, The Hague, 1975).
12. Collard, M. & Aiello, L. C. *Nature* **404**, 339–340 (2000).

Richmond and Strait argue that evidence of ancestral knuckle-walking is retained in the distal radius of *Australopithecus*¹, basing their conclusions on a canonical variate analysis that uses four measurements of the distal radial joint surface in higher primates. We dispute their claim that early hominids retained a specialized wrist morphology associated with knuckle-walking and question the biological and mechanical interpretations they use to attribute knuckle-walking features to the radius of *Australopithecus*.

The authors believe that in African apes there is a mechanism that 'locks' the wrist into place, which they base on their inference that the scaphoid conforms with a distally facing notch in the dorsal joint margin of the radius in full extension. They suggest that this 'locking mechanism' acts by limiting wrist extension and argue that it exists

because chimpanzees do not consistently recruit the major flexor muscles of the digits during knuckle-walking.

We question these conclusions on several grounds. First, very little motion occurs at the radioscapoid joint — Richmond and Strait's X-ray evidence shows that it takes place almost entirely at the midcarpal joint. All synovial joints conform throughout their range of motion, and would be expected to do so in full flexion. Most important, synovial joint stability derives from its muscles and ligaments², not from its essentially frictionless joint surfaces: the primary role of joint-surface geometry is to assure that the joint reaction force remains orthogonal to rigid-body contact surfaces, and that velocity vectors remain tangential throughout the range of motion³.

Neither do normal joints lock. Locking would prevent motion and thus the performance of cartilage-sparing, negative work by the joint's soft-tissue envelope. Cartilage congruity maintains joint reaction force normality throughout the flexion–extension cycle, especially under high external loads. Neither peripheral nor edge loading occurs in normal, stable joints³. Further, neither the location nor direction of the joint reaction force can be determined from an X-ray image — comprehensive free-body analysis is required.

Finally, tendons and fascial sheaths are crucial to joint stability, particularly those of the long digital flexors⁴. An absence of electromyographic activity suggests that passive recruitment of the joint capsule and the muscle's connective-tissue envelope are all that is required for joint stability during light-to-moderate loading⁵, a possibility considered as likely in a study⁶ cited by Richmond and Strait. It is unclear how other joints of the wrist and hand might be sustained (see the authors' Fig. 1b) during significant loading if no tendons are recruited.

This confusion extends to ontogeny. Details of surface conformity, as in Richmond and Strait's 'scaphoid notch', result from cartilage modelling (they cannot be determined by hereditary 'descriptive' specification⁷). Chondral modelling modulates the curvature, shape, congruence and smoothness of a joint's articular surface⁸ and postnatal joint motion and joint reaction forces determine the final adult joint contours^{9,10}. It is therefore unlikely that *Australopithecus* retained details of joint contours specific to knuckle-walking without performing that behaviour.

Richmond and Strait used four variables for their canonical variate analysis, of which two were the breadth and orientation of their 'scaphoid notch' — defined as the length of the portion along the dorsal margin of the distal radial articular facet that is bevelled. This 'notch' must be either a special morphological substructure of the joint

margin, or simply some fraction of it. It cannot be the former, as the authors' data are neither dichotomous nor bimodal (or all primates would have scaphoid notches and would, by induction, be adapted to knuckle-walking). Their notch therefore appears to be an undefined (neither arc nor chord location is specified) estimate of the curvature of some fraction of this already diminutive joint margin whose form is dictated by chondral modelling. Without the first axis, *Homo* and *Pongo* collapse into a single distribution, as do terrestrial monkeys and the African apes, so this analysis cannot distinguish their highly distinctive locomotor patterns, and all but one hominid (the extensively damaged KNM-ER-20419) would distribute with *Homo*.

Human and chimpanzee distal radii do differ. However, their dissimilarity derives primarily from a difference in angulation between the epiphysis and diaphysis (see www.kent.edu/anthropology/naturefigure). The greater angulation in African apes is the source of another character that Richmond and Strait refer to as distal projection. As the entire epiphysis is so angled, such differences are more likely to be due to morphogen or paracrine signalling during pattern formation, for example by modulating the parathyroid-hormone-related protein-Indian Hedgehog loop^{11,12}, and to asymmetric physis compression induced by knuckle-walking during ontogeny (the same processes that produce human knee valgus¹³).

C. Owen Lovejoy*†, Kingsbury G. Heiple†, Richard S. Meindl*

*Division of Biomedical Sciences, Department of Anthropology, Kent State University, Kent, Ohio 44242, USA

†Division of Surgery, Department of Orthopaedics, School of Medicine, Case Western Reserve University, Cleveland, Ohio 44206, USA
e-mail: olovejoy@aol.com

- Richmond, B. G. & Strait, D. S. *Nature* **404**, 382–385 (2000).
- Martin, R. B., Burr, D. B. & Sharkey, N. A. *Skeletal Tissue Mechanics* (Springer, New York, 1998).
- Burstein, A. H. & Wright, T. M. *Fundamentals of Orthopaedic Biomechanics* (Williams & Wilkins, Baltimore, 1994).
- Shadwick, E. D. *J. Appl. Physiol.* **68**, 1033–1040 (1990).
- Roberts, T. J. *et al. Science* **275**, 1113–1115 (1997).
- Susman, R. L. & Stern, J. T. *Jr Am. J. Phys. Anthropol.* **50**, 565–574 (1979).
- Thorogood, P. in *Embryos, Genes, and Birth Defects* (ed. Thorogood, P.) 1–16 (Wiley, Chichester, 1997).
- Frost, H. M. *Anat. Rec.* **255**, 162–174 (1999).
- Ogden, J. A. in *Fundamental and Clinical Bone Physiology* (ed. Urist, M. R.) 108–171 (Lippincott, Philadelphia, 1980).
- Hamrick, M. J. *Theor. Biol.* **201**, 201–208 (1999).
- Karp, S. J. *et al. Development* **127**, 543–548 (2000).
- St-Jacques, B., Hammerschmidt, M. & McMahon, A. P. *Genes Dev.* **13**, 2072–2086 (1999).
- Lovejoy, C. O., Cohn, M. J. & White, T. W. *Proc. Natl Acad. Sci. USA* **96**, 13247–13252 (1999).

Richmond and Strait reply — Dainton's argument that knuckle-walking may have evolved in parallel in *Gorilla* and the *Pan/Homo* ancestor is possible, but unparsimonious. Moreover, patterns of African ape carpal growth¹ do not provide an

adequate test of knuckle-walking homology, given evidence that development does not always faithfully reflect homology². Siangmangs and gibbons differ in skeletal growth³, but no one disputes the homology of their suspensory specializations.

Regarding Dainton's points about our functional interpretation, an obtuse angulation between the scaphoid and lunate facets resists compressive stresses as part of a complex that stabilizes the wrist in extension during knuckle-walking⁴. With respect to the statistics, our cluster analysis (Fig. 2c), the Mahalanobis D^2 distances, and posterior probabilities of group assignment show that *A. afarensis* and *A. anamensis* radii are more similar to *Gorilla* and *Pan* than either are to *Pongo*. Finally, although knuckle-walking traits may have arisen to serve extension-limiting functions in other forms of locomotion, there is no evidence to support this. Our hypothesis meets Lauder's definition of a "best-case scenario"⁵, in which the extinct taxon whose behaviour is being reconstructed lies within a clade that includes extant members for which a form is exclusively related to a function.

Lovejoy *et al.* misinterpret our discussion of radio-scaphoid joint function. All the anthropoids we examined have scaphoid notches. However, the morphology of the notch, in combination with features such as distal projection, distinguishes knuckle-walkers from other taxa (see www.anthro.uiuc.edu/faculty/richmond/hominidradii.html). African apes differ from other anthropoids in that the radius and scaphoid conform posteriorly after minimal extension, such that "no further movement is possible without disengaging the articular surfaces ... this part of the joint achieves a 'locked' or 'close-packed' position"⁴. Such terminology is conventional⁶. Muscle and other tissues can play a role in promoting joint stability, and we explicitly cited experimental evidence on the subject. However, the digital flexors are not recruited during knuckle-walking at slow-to-moderate speeds. Active recruitment is crucial in maintaining tension to enable the tendons to store and return elastic energy⁷. Moreover, passive tension of the digital flexors is unlikely to be significant during knuckle-walking because the fingers are flexed. Regardless, active or passive muscle tension across the wrist does not necessarily diminish the importance of joint morphology in providing stability.

Lovejoy *et al.* suggest developmental hypotheses to explain variation in joint morphology. We disagree with their implication that species-specific aspects of joint shape have little hereditary basis. During early development, joint morphology is heavily influenced by genetic factors, to the extent that normal articular topography will appear even if the adjacent surface is

absent or if limb paralysis is induced⁸. Although the employment of abnormal positional behaviours influences articular development⁹, it is unclear how, or whether, the absence of specific behaviours influences joint growth. Current evidence suggests that bone length and aspects of joint shape are not heavily influenced by physiological variations in activity during growth, unlike trabecular structure, and cortical bone thickness, density and curvature¹⁰. Thus, functionally sensitive features such as femoral bicondylar angle and phalangeal curvature suggest both bipedalism and arboreality in *Australopithecus*, whereas radiocarpal joint shape might be a product of evolutionary history.

Although Lovejoy *et al.* reject the possibility that functionally relevant traits might be retained through phylogenetic inertia, they have argued¹¹, in reference to australopithecine joint morphology, that one cannot "expect an animal that has recently evolved bipedality to no longer have any anatomical characters that are not also found in modern-day apes". This thesis is central to their^{11,12} long-held argument that early australopithecines are obligate bipeds but retain arboreal features without practising arboreality. It might be concluded from Lovejoy *et al.*'s present developmental argument that these very same fossil hominids are knuckle-walkers. We doubt they believe this, and neither do we. Phylogenetic inertia remains the most plausible interpretation of the morphology we identified. How functionally to interpret primitive retentions is the central problem in reconstructing behaviour in fossil hominids. We still do not fully understand the genetic and epigenetic influences on bone development, and need to identify bone morphology that provides a faithful record of an individual's biomechanical activity.

Brian G. Richmond*†, David S. Strait†‡

*Department of Anthropology, University of Illinois, Urbana, Illinois 61820, USA

email: brich@uiuc.edu

†Department of Anthropology, George Washington University, Washington DC 20052, USA

‡Department of Anatomy, New York College of Osteopathic Medicine, Old Westbury, New York 11949, USA

- Dainton, M. & Macho, G. A. *J. Hum. Evol.* **36**, 171–194 (1999).
- Hall, B. K. (ed.) *Homology: The Hierarchical Basis of Comparative Biology* (Academic, San Diego, CA, 1994).
- Junger, W. L. & Cole, M. S. *J. Hum. Evol.* **23**, 93–105 (1992).
- Jenkins, F. A. J. & Fleagle, J. G. in *Primate Functional Morphology and Evolution* (ed. Tuttle, R. H.) 213–231 (Mouton, The Hague, 1975).
- Lauder, G. V. in *Functional Morphology in Vertebrate Paleontology* (ed. Thompson, J. J.) 1–18 (Cambridge Univ. Press, Cambridge, 1995).
- Tuttle, R. H. *Am. J. Phys. Anthropol.* **26**, 171–206 (1967).
- Roberts, T. J. *et al. Science* **275**, 1113–1115 (1997).
- Hamrick, M. W. *J. Theor. Biol.* **201**, 201–208 (1999).
- Nakatsukasa, M. *et al. Folia Primatol.* **64**, 1–29 (1995).
- Lanyon, L. E. *J. Zool.* **192**, 457–466 (1980).
- Oliwenstein, L. *Science* **269**, 476–477 (1995).
- Latimer, B. M. & Lovejoy, C. O. *Am. J. Phys. Anthropol.* **82**, 125–133 (1990).

Guard cell abscisic acid signalling and engineering drought hardiness in plants

Julian I. Schroeder, June M. Kwak & Gethyn J. Allen

Cell and Developmental Biology Section, Division of Biology and Center for Molecular Genetics, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0116, USA

Guard cells are located in the epidermis of plant leaves, and in pairs surround stomatal pores. These control both the influx of CO₂ as a raw material for photosynthesis and water loss from plants through transpiration to the atmosphere. Guard cells have become a highly developed system for dissecting early signal transduction mechanisms in plants. In response to drought, plants synthesize the hormone abscisic acid, which triggers closing of stomata, thus reducing water loss. Recently, central regulators of guard cell abscisic acid signalling have been discovered. The molecular understanding of the guard cell signal transduction network opens possibilities for engineering stomatal responses to control CO₂ intake and plant water loss.

Desiccation of crops and horticultural plants during periods of drought causes severe and often irreversible damage and lost yields. Fresh water scarcity has been identified as one of the principal global problems in the new century¹, and plants account for around 65% of global fresh water use. Engineering of the stomatal guard cell signal transduction network could provide a major contribution to more sustainable water use during droughts.

Plants are rooted in one place and therefore need to respond effectively to environmental stresses. Intricate signalling pathways form networks that regulate plant responses to the environment. However, many of the principal signalling components and their interactions remain largely unknown.

Guard cell signalling integrates water status, hormonal stimuli, light, CO₂ levels and other environmental conditions to regulate stomatal apertures. Guard cells are a well-developed model system for understanding how components interact within a signalling network in a single cell. They are well suited for dissecting the functions of genes and proteins in signalling cascades. Consequently, powerful techniques have been developed for guard cell signalling studies, enabling combined molecular genetic, cell biological, biophysical and proteomic and genomic analyses.

Guard cells respond at the single cell level to physiological stimulation, allowing cell biological analyses in response to diverse stimuli. Guard cells respond to most of the classical plant hormones, which illustrates that unidentified receptors and early signalling mechanisms function in these cells.

A network of ion channels in the plasma membrane and the vacuolar membrane of guard cells, which controls stomatal movements, has been characterized (see Box 1). These ion channels are targets of early signalling branches and provide probes to identify and characterize upstream transducers.

Abscisic acid (ABA)-induced stomatal closing is mediated by a reduction in the turgour pressure of guard cells, which requires an efflux of potassium and anions, sucrose removal and the conversion of malate to osmotically inactive starch². Box 1 shows an extension of early models for the proposed functions of ion channels during ABA-induced stomatal closing^{3,4}.

Reconstitution in *Xenopus* oocytes

Cloning of an ABA receptor has remained a goal of plant cell signalling research and plant stress tolerance engineering. Heterologous expression systems, such as *Xenopus laevis* oocytes, have been used to clone animal cell receptors, and this technique has been used in attempts to identify ABA receptors and transduction

mechanisms^{5,6}. Simultaneous injection of guard-cell-derived poly(A⁺) RNA with RNA of the *Arabidopsis* inward-rectifying K⁺ channel *KAT1* reconstitutes ABA-induced downregulation of KAT1-mediated K_{in}⁺ channel currents (Box 1a) in oocytes⁶.

ABA causes activation of endogenous Ca²⁺-activated Cl⁻ currents in oocytes injected with messenger RNA from drought-stressed tobacco leaves⁵. By fractionation of the complementary DNA library, *Nt-Syr1*, a homologue of human and yeast syntaxins, was isolated. However, *Nt-Syr1* expression in oocytes constitutively activates Cl⁻ currents in oocytes without ABA addition⁵. Further work is required to determine whether *Nt-Syr1* is actually involved in the heterologous ABA response in oocytes.

Syntaxins and SNARE proteins have an integral function in vesicle trafficking in animal, yeast and plant cells. In tobacco guard cells, ABA regulation of slow-activating sustained (S-type) anion, K_{out}⁺ and K_{in}⁺ channels (Box 1a) is inhibited by microinjection of truncated *Nt-Syr1* or injection of the neurotoxin *Clostridium botulinum* C, which inhibits syntaxin-dependent vesicle trafficking⁵. During stomatal closing, the plasma membrane surface area of guard cells decreases to accommodate volume reductions⁷, which may require syntaxins. On the basis of the guard cell ion channel responses, syntaxins were proposed to function not only in membrane vesicle trafficking, but also in ABA signalling⁵ (Fig. 1). These studies show functional expression of ABA perception and transduction mechanisms in oocytes, which may therefore be suitable for identifying ABA receptors.

Plasma membrane Ca²⁺ channels

Calcium influx and cytoplasmic calcium ([Ca²⁺]_{cyt}) increases are important for guard cell ABA signal transduction (Box 1)^{4,8,9}. Impairment of ABA-induced [Ca²⁺]_{cyt} increases in the ABA-insensitive *Arabidopsis* mutants *abi1-1* and *abi2-1* provides genetic evidence for the importance of [Ca²⁺]_{cyt} elevations in ABA signalling¹⁰ (Fig. 1). In *Vicia* guard cells, Ca²⁺ influx is enhanced when the membrane potential is hyperpolarized below about -120 mV^{11,12}, but ABA shifts the threshold for influx to potentials above -80 mV, which suggests that voltage-dependent Ca²⁺ influx and ABA signalling are closely linked¹². Moreover, the probability of ABA-induced [Ca²⁺]_{cyt} elevations is increased if guard cells are maintained in hyperpolarizing conditions¹²⁻¹⁴.

The upstream mechanisms by which ABA activates guard cell plasma membrane Ca²⁺ channels had remained unknown. Reactive oxygen species (ROS) were recently identified as an activator of hyperpolarization-activated Ca²⁺-permeable channels (*I*_{Ca}) in *Arabidopsis* guard cells¹⁵. ROS and *I*_{Ca} were established as com-

ponents contributing to early ABA signalling because ABA causes a rapid increase in ROS and I_{Ca} activation in guard cells¹⁵, whereas ROS-induced stomatal closure and I_{Ca} activation are abolished in the ABA-insensitive recessive mutant *gca2* (refs 15, 16) (Fig. 1).

A hyperpolarization-activated channel similar to I_{Ca} was identified in *Vicia* guard cells¹⁷. This channel shows bursts of increased activity in response to cytoplasmic ABA in isolated membrane patches¹⁷. ABA activation of these Ca^{2+} channels shows repetitive transients and rapidly desensitizes, suggesting that a second messenger might be generated locally at the plasma membrane in response to ABA. The Ca^{2+} influx channel is itself downregulated by $[Ca^{2+}]_{cyt}$ elevation¹⁷, which may be part of a regulating system that generates complex ABA-induced Ca^{2+} oscillations (Fig. 2).

I_{Ca} channels are activated by ROS¹⁵ and so could act as a focal point for integrating other, non-ABA, stress signals and the oxidative state of guard cells. For example, pathogenic elicitors cause ROS production in guard cells and stomatal pore closure¹⁸. Furthermore, ozone, which causes oxidative stress, downregulates K_{in}^+ channels and inhibits stomatal opening¹⁹, raising the possibility that ozone regulates I_{Ca} -type channels.

Ca^{2+} release and oscillations

In *Commelina* guard cells, plasma membrane Ca^{2+} influx is more strongly activated at high ABA concentrations, whereas at low ABA concentrations the InsP3 and cADPR pathways (see Box 1) predominate^{20,21}. Combined application of the phospholipase C (PLC) inhibitor U-73122 and the cADPR inhibitor nicotinamide inhibit ABA-induced K^+ efflux and stomatal closing^{21,22}, whereas each inhibitor alone produces only weak inhibition^{13,21,22}, suggesting that these two Ca^{2+} -release signalling pathways operate in parallel during ABA signalling.

Cytosolic Ca^{2+} oscillations (Fig. 2) have been proposed as necessary for stomatal closure⁹. The four stomatal closing stimuli

ABA, ROS, low temperature and external Ca^{2+} treatment produce Ca^{2+} oscillations in wild-type *Arabidopsis* guard cells^{13,23} (Fig. 2). In contrast, in the *Arabidopsis* V-type H^+ -ATPase mutant *det3*, ROS and external Ca^{2+} generate prolonged $[Ca^{2+}]_{cyt}$ elevations (plateaux) in guard cells, and long-term stomatal closure by these stimuli is inhibited²³. Experimentally re-establishing external Ca^{2+} -induced Ca^{2+} oscillations in *det3* guard cells by repetitive hyperpolarizations rescues stomatal closure in *det3*. Furthermore, stomatal closure is abolished in wild-type guard cells when non-oscillating Ca^{2+} plateaux are experimentally imposed²³. Overall, these results suggest that Ca^{2+} oscillations encode information required for stomatal closure. New techniques that combine Ca^{2+} indicators (Fig. 2) and molecular genetics in *Arabidopsis* guard cells^{10,14,15,23} provide a powerful combination for future analyses of $[Ca^{2+}]_{cyt}$ oscillation generation and downstream processing in plants.

More lipid-derived second messengers

ABA stimulates production of another inositol-phosphate second messenger, myo-inositol-hexakisphosphate (InsP6), to a greater extent than InsP3 in guard cells²⁴ (Fig. 1). InsP6 inhibits K_{in}^+ channels in a Ca^{2+} -dependent manner, and more efficiently than InsP3 (ref. 24). It was hypothesized that microinjection of InsP3 functions by affecting the InsP6 pathway²⁴. Whether InsP6 causes $[Ca^{2+}]_{cyt}$ elevations and whether these two messengers function in the same or separate signalling branches is unknown.

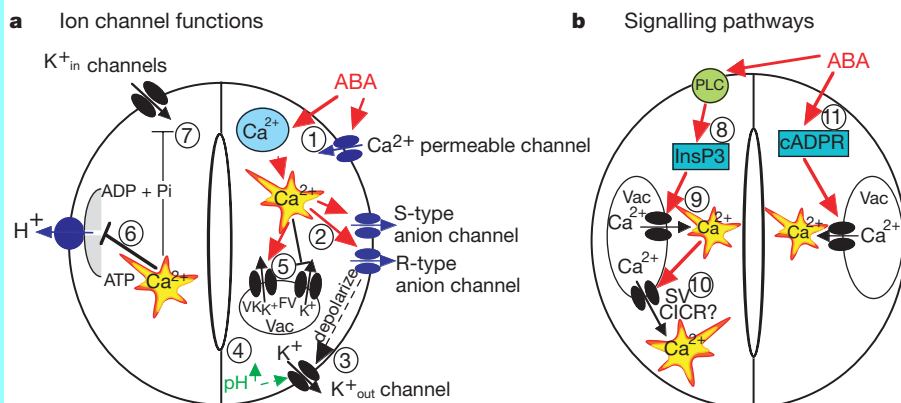
Another phospholipase acts in ABA signalling. The level of phosphatidic acid (PtdOH) generated by phospholipase D (PLD) transiently increases after ABA treatment of aleurone cells and guard cells²². PtdOH promotes partial (50%) stomatal closure, inhibits stomatal opening and partially inactivates K_{in}^+ channel currents. PtdOH does not elicit a $[Ca^{2+}]_{cyt}$ increase in guard cells²² and PLD is therefore likely to act downstream of or parallel to ABA-induced $[Ca^{2+}]_{cyt}$ elevations (Fig. 1).

Box 1

Components and pathways of guard cell ABA signalling

a, ABA is detected by as yet unidentified receptors (right guard cell) and induces cytosolic Ca^{2+} elevations (1) through extracellular Ca^{2+} influx and release from intracellular stores (for reviews see refs 2, 8, 9). (2) $[Ca^{2+}]_{cyt}$ elevations activate two types of anion channel that mediate anion release from guard cells, slow-activating sustained (S-type) or rapid transient (R-type) anion channels^{3,8}. (3) Anion efflux causes depolarization, which activates outward-rectifying K^+ (K_{out}^+) channels and results in K^+ efflux from guard cells^{2,3}. (4) ABA causes an alkalization of the guard cell cytosol, which enhances K_{out}^+ channel activity⁴⁴. Overall, the long-term efflux of both anions and K^+ from guard cells contributes to the loss of guard cell turgor, leading to stomatal closing³. Over 90% of the ions released from the cell during stomatal closing must be first released from vacuoles into the cytosol. (5) At the vacuole, $[Ca^{2+}]_{cyt}$ elevation activates vacuolar K^+ (VK) channels, which are thought to mediate Ca^{2+} -induced K^+ release from the vacuole⁸. In addition, fast vacuolar (FV) channels can mediate K^+ efflux from guard cell vacuoles at resting $[Ca^{2+}]_{cyt}$ ⁴⁵. ABA also inhibits ion uptake, which is required for stomatal opening (left guard cell). (6) $[Ca^{2+}]_{cyt}$ elevations inhibit the electrogenic plasma membrane proton-extruding H^+ -ATPases⁴⁶ and K^+ uptake (K_{in}^+) channels^{2,3,44}. (7). Initiation of ion efflux (1–5, right guard cell) and inhibition of stomatal opening processes (6, 7, left guard cell) provide a mechanistic basis for ABA-induced stomatal closing⁸.

b, Ca^{2+} -releasing second messengers that mediate ABA-induced stomatal closing. Second messengers have been implicated in ABA signalling in guard cells, including InsP3 and cADPR. (8) Biochemical evidence indicates that ABA stimulates InsP3 production in guard cells². (9) InsP3 can release Ca^{2+} from intracellular stores, inhibit K_{in}^+ channels and cause stomatal closure^{9,13,22,44}. (10) $[Ca^{2+}]_{cyt}$ elevations may be amplified by a Ca^{2+} -induced Ca^{2+} release (CICR) mechanism from the vacuole through activation of the Ca^{2+} -dependent slow vacuolar (SV) channel^{8,45}. Data questioning this SV model for CICR⁴⁷ and recent other data supporting it⁴⁸ indicate the potential for molecular genetic analyses. (11) ABA stimulates the production of cADPR in plant cells⁴⁹, which also increases $[Ca^{2+}]_{cyt}$ in guard cells^{20,45} and promotes stomatal closing.



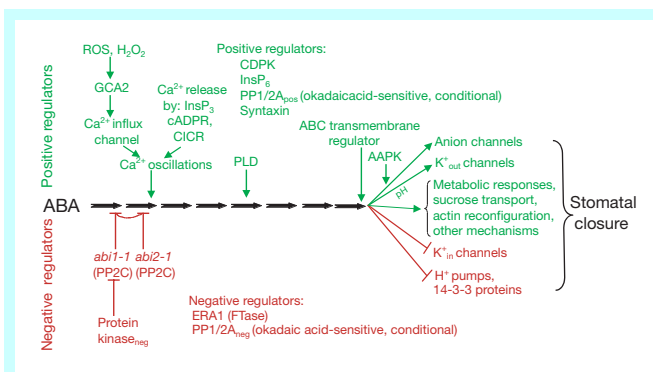


Figure 1 Simplified model for proposed functions of positive and negative regulators mediating guard cell ABA signal transduction. Positive ABA-activated regulators (top, in green) and negative ABA-signalling regulators (bottom, in red) in guard cells are shown. The sequence of events among regulators remains largely unknown and will require in-depth analyses. Positive and negative regulators for which the position relative to cytosolic Ca^{2+} elevations remains to be determined are listed alphabetically. Some parallel signalling branches are excluded for simplicity.

Protein kinases and phosphatases

Phosphorylation events are central mediators of ABA signalling in guard cells^{25–31}. Some of the underlying protein kinase³¹ and phosphatase genes^{16,32} have been cloned. Both positively and negatively regulating protein kinases, and positively and negatively regulating PP2C- and PP1/2A-type protein phosphatases (Fig. 1, PP2Cs, PP1/2A_{pos} and PP1/PP2A_{neg}) have been proposed to function in ABA signalling (for details see refs 10, 16, 29, 32). ABA signalling can be transduced by several types of positively regulating protein kinases^{31,33,34}. For example, biochemical analyses have led to the hypothesis that Ca^{2+} -dependent protein kinases (CDPKs) may function upstream of ABA-activated Ca^{2+} -independent kinases in maize protoplasts and *Vicia* guard cells^{28,34}.

In-gel phosphorylation assays showed that in guard cells ABA rapidly activates a Ca^{2+} -independent protein kinase^{27,28} with a relative molecular mass of 48,000. A corresponding guard-cell-specific protein kinase, AAPK, was isolated by two-dimensional gel separation and mass spectrometry³¹. Transient transformation of guard cells with a dominant mutant form of AAPK with reduced activity impairs ABA activation of S-type anion channels and disrupts ABA-induced stomatal closing³¹ (Fig. 1).

With one exception³⁰, dominant ABA-insensitive mutations in the known protein kinase and phosphatase genes *ABI1*, *ABI2* and *AAPK*, have been analysed in guard-cell signalling^{25,29,31}. It will be important to isolate recessive loss-of-function mutations in these enzymes^{16,30} to test the proposed functions of the wild-type proteins (Fig. 1). Scaffolding mechanisms may hold the key for target specificity of the many proposed protein kinases and phosphatases (Fig. 1).

Scaffolding mechanisms and signalling complexes

In yeast and animal cells proteins are clustered into signalling complexes or 'transducisomes', which provide a structural basis for specificity in signalling³⁵. Little is known about signal transduction complexes in plants, although more than ten genes in the *Arabidopsis* genome encode for 14-3-3 proteins, which function as scaffolding proteins by linking protein kinases to their targets or by mediating protein–protein interactions³⁶. 14-3-3 proteins have a direct function in blue-light-induced protein kinase activation of guard cell plasma membrane H^+ -ATPases, by controlling phosphorylation of serine and threonine residues in the carboxy-terminus of the ATPases³⁷.

There must also be additional mechanisms for forming signalling complexes in guard cells. Pharmacological studies have suggested

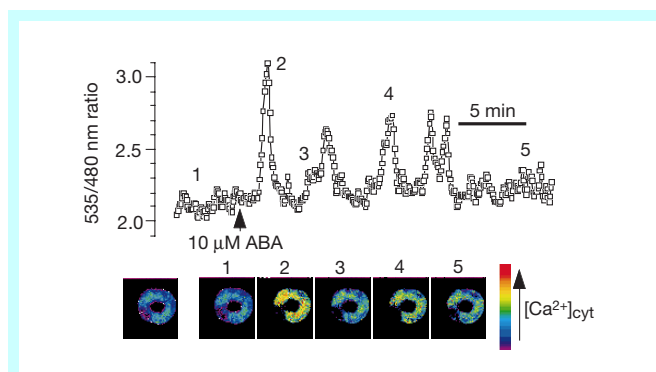


Figure 2 ABA-induced $[\text{Ca}^{2+}]_{\text{cyt}}$ oscillations in *Arabidopsis* guard cells expressing yellow cameleon 2.1. $[\text{Ca}^{2+}]_{\text{cyt}}$ oscillations elicited by ABA are indicated by an increase in the fluorescence emission ratio 535/480 nm¹⁴. Bottom, pseudo-coloured ratio images of cameleon fluorescence in a stomate after ABA application. The trace is from one of the imaged guard cells; images are shown at the points marked.

that an ATP-binding-cassette transmembrane regulator (ABC protein) may form a junction point in guard cell ABA signalling, thus simultaneously regulating S-type anion channels and K^+ _{out} channels³⁸ (Fig. 1). Inhibitors of sulphonylurea receptor (SUR) ABC proteins, such as glibenclamide, inhibit both S-type anion channels and K^+ _{out} channels, and abolish stomatal closure triggered by ABA or external Ca^{2+} . The SUR activator cromakalim reverses the glibenclamide inhibition of these channels^{38,39}.

Several soluble signalling proteins have been proposed to regulate guard cell ion channels (Fig. 1). Farnesyltransferases mediate membrane targeting of specific soluble signalling proteins by attachment of a hydrophobic farnesyl group to C-terminal target sequences. The ERA1 farnesyltransferase β -subunit is a negative regulator of ABA signal transduction in seeds⁴⁰ and guard cells⁴¹, which suggests farnesylation-dependent membrane targeting of soluble, negative regulators of ABA signalling (Fig. 1). Future proteomic-based research^{31,42} should provide a more complete molecular characterization of transducisomes in guard cells.

Molecular engineering

It would be beneficial to agriculture for crop plants to show wide stomatal opening for CO_2 intake when water is available, but to close stomata during drought periods, thereby slowing desiccation and damage. However, conventional high-yield breeding approaches may have contributed to selection of crop plants with reduced stomatal ABA responsiveness, because genes controlling guard cell signalling are also expressed in other tissues and control other yield parameters (for example, refs 15, 16, 29, 32, 40, 41).

Studies on stomatal ion channel regulation have identified mechanisms that enhance ABA-induced stomatal closing. Deletion of the *ERA1* gene causes ABA hypersensitivity of both anion channel activation and stomatal closing, and during drought treatment *era1* plants show reduced wilting and enhanced stomatal closure⁴¹. Notably, when *era1* plants are watered the stomatal pores open, although to a lesser extent than in the wild type, allowing CO_2 influx and growth. Intragenic suppressor mutants of *abi1-1* (*abi1-1R*) also show reduced water loss during drought³⁰; however, *era1* and *abi1-1R* alleles have additional growth and developmental phenotypes.

Based on the model in Fig. 1, downregulation of negative regulators (such as ERA1) or upregulation of positive regulators specifically in guard cells using guard-cell-specific⁴³ and stress-responsive promoters, or isolation of guard cell specific mutants, will allow engineering of plants that lose less water during drought. Some mutations may constitutively reduce stomatal opening, which could reduce the water requirements of horticultural plants and turf grass, or in agricultural regions of marginal fresh water availability¹.

For plants growing in humid regions, weakening of ABA signal transduction components (Fig. 1) may enhance stomatal opening and CO₂ intake for carbon fixation and growth³¹.

Targeted mutations that only affect stomatal movements will also provide powerful tools to quantitatively analyse photosynthetic carbon fixation, gas exchange and plant growth as a direct function of specific alteration of stomatal apertures. Such work, together with engineering of leaf waxiness and osmoprotectant production, may also contribute to future molecular engineering of improved plant survival during droughts.

Recent efforts have led to major advances in understanding the molecular mechanisms of guard cell ABA signalling. A model for the functions of ion channels in regulating stomatal movements (Box 1) has provided a framework for the characterization of new, early ABA transduction mechanisms (Fig. 1). However, many components within the nonlinear and branched pathways of this network remain to be identified. Moreover, the relative positioning of the mechanisms identified and the exact functions of pharmacologically predicted transducers remain to be determined using combined genetic, cell biological and biophysical approaches. Given the large gene families, gene duplications and redundancies in the *Arabidopsis* genome, it will be important to identify guard-cell-expressed isoforms of these transducers using DNA microarrays, and to use guard cell promoters⁴³ to cell-specifically manipulate redundant, and often widely expressed, genes. Guard cell and other single cell analyses of signal transduction are moving towards achieving a time-resolved understanding of an entire plant cell signalling network. Postgenomic and proteomic approaches will be essential for future advances. The new insights into guard cell and ABA signalling, combined with interdisciplinary approaches, will lead to future revelations in this field that will be relevant to plant signal transduction, environmental ecology and agriculture. □

- Postel, S. L., Daily, G. C. & Ehrlich, P. R. Human appropriation of renewable fresh water. *Science* **271**, 785–788 (1996).
- MacRobbie, E. A. C. Signal transduction and ion channels in guard cells. *Phil. Trans. R. Soc. Lond. B* **353**, 1475–1488 (1998).
- Schroeder, J. I. & Hedrich, R. Involvement of ion channels and active transport in osmoregulation and signaling of higher plant cells. *Trends Biochem. Sci.* **14**, 187–192 (1989).
- McAinsh, M. R., Brownlee, C. & Hetherington, A. M. Absciscic acid-induced elevation of guard cell cytosolic Ca²⁺ precedes stomatal closure. *Nature* **343**, 186–188 (1990).
- Leyman, B., Geelen, D., Quintero, F. J. & Blatt, M. R. A tobacco syntaxin with a role in hormonal control of guard cell ion channels. *Science* **283**, 537–540 (1999).
- Sutton, F., Paul, S. S., Wang, X. -Q. & Assmann, S. M. Distinct absciscic acid signaling pathways for modulation of guard cell versus mesophyll cell potassium channels revealed by expression studies in *Xenopus laevis* oocytes. *Plant Physiol.* **124**, 223–230 (2000).
- Homann, U. & Thiel, G. Unitary exocytotic and endocytotic events in guard-cell protoplasts during osmotically driven volume changes. *FEBS Lett.* **460**, 495–499 (1999).
- Ward, J. M., Pei, Z. -M. & Schroeder, J. I. Roles of ion channels in initiation of signal transduction in higher plants. *Plant Cell* **7**, 833–844 (1995).
- McAinsh, M. R., Brownlee, C. & Hetherington, A. M. Calcium ions as second messengers in guard cell signal transduction. *Physiol. Planta* **100**, 16–29 (1997).
- Allen, G. J., Kuchitsu, K., Chu, S. P., Murata, Y. & Schroeder, J. I. *Arabidopsis* *abi1-1* and *abi2-1* phosphatase mutations reduce absciscic acid-induced cytoplasmic calcium rises in guard cells. *Plant Cell* **11**, 1785–1798 (1999).
- Gilroy, S., Fricker, M. D., Read, N. D. & Trewavas, A. J. Role of calcium in signal transduction of *Commelina* guard cells. *Plant Cell* **3**, 333–344 (1991).
- Grabov, A. & Blatt, M. R. Membrane voltage initiates Ca²⁺ waves and potentiates Ca²⁺ increases with absciscic acid in stomatal guard cells. *Proc. Natl Acad. Sci. USA* **95**, 4778–4783 (1998).
- Staxen, I. *et al.* Absciscic acid induces oscillations in guard-cell cytosolic free calcium that involve phosphoinositide-specific phospholipase C. *Proc. Natl Acad. Sci. USA* **96**, 1779–1784 (1999).
- Allen, G. J. *et al.* Cameleon calcium indicator reports cytoplasmic calcium dynamics in *Arabidopsis* guard cells. *Plant J.* **19**, 735–747 (1999).
- Pei, Z. -M. *et al.* Calcium channels activated by hydrogen peroxide mediate absciscic acid signalling in guard cells. *Nature* **406**, 731–734 (2000).
- Himmelbach, A., Iten, M. & Grill, E. Signalling of absciscic acid to regulate plant growth. *Phil. Trans. R. Soc. Lond. B* **353**, 1439–1444 (1998).
- Hamilton, D. W. A., Hills, A., Kohler, B. & Blatt, M. R. Ca²⁺ channels at the plasma membrane of stomatal guard cells are activated by hyperpolarization and absciscic acid. *Proc. Natl Acad. Sci. USA* **97**, 4967–4972 (2000).
- Lee, S. *et al.* Oligogalacturonic acid and chitosan reduce stomatal aperture by inducing the evolution of reactive oxygen species from guard cells of tomato and *Commelina communis*. *Plant Physiol.* **121**, 147–152 (1999).
- Torsethagen, G., Pell, E. J. & Assmann, S. M. Ozone inhibits guard cell K⁺ channels implicated in stomatal opening. *Proc. Natl Acad. Sci. USA* **96**, 13577–13582 (1999).
- Leckie, C. P., McAinsh, M. R., Allen, G. J., Sanders, D. & Hetherington, A. M. Absciscic acid-induced stomatal closure mediated by cyclic ADP-ribose. *Proc. Natl Acad. Sci. USA* **95**, 15837–15842 (1998).
- MacRobbie, E. A. C. ABA activates multiple Ca²⁺ fluxes in stomatal guard cells, triggering vacuolar K⁺ (Rb⁺) release. *Proc. Natl Acad. Sci. USA* **97**, 12361–12368 (2000).
- Jacob, T., Ritchie, S., Assmann, S. M. & Gilroy, S. Absciscic acid signal transduction in guard cells is mediated by phospholipase D activity. *Proc. Natl Acad. Sci. USA* **96**, 12192–12197 (1999).
- Allen, G. J. *et al.* Alteration of stimulus-specific guard cell calcium oscillations and stomatal closing in *Arabidopsis* *det3* mutant. *Science* **289**, 2338–2342 (2000).
- Lemtiri-Chlieh, F., MacRobbie, E. A. C. & Brealey, C. A. Inositol hexakisphosphate is a physiological signal regulating the K⁺-inward rectifying conductance in guard cells. *Proc. Natl Acad. Sci. USA* **97**, 8687–8692 (2000).
- Armstrong, F. *et al.* Sensitivity to absciscic acid of guard cell K⁺ channels is suppressed by *ABI1-1*, a mutant *Arabidopsis* gene encoding a putative protein phosphatase. *Proc. Natl Acad. Sci. USA* **92**, 9520–9524 (1995).
- Schmidt, C., Schelle, I., Liao, Y. J. & Schroeder, J. I. Strong regulation of slow anion channels and absciscic acid signaling in guard cells by phosphorylation and dephosphorylation events. *Proc. Natl Acad. Sci. USA* **92**, 9535–9539 (1995).
- Li, J. & Assmann, S. An absciscic acid-activated and calcium-independent protein kinase from guard cells of fava bean. *Plant Cell* **8**, 2359–2368 (1996).
- Mori, I. & Muto, S. Absciscic acid activates a 48-kilodalton protein kinase in guard cell protoplasts. *Plant Physiol.* **113**, 833–840 (1997).
- Pei, Z. -M., Kuchitsu, K., Ward, J. M., Schwarz, M. & Schroeder, J. I. Differential absciscic acid regulation of guard cell slow anion channels in *Arabidopsis* wild-type and *abi1* and *abi2* mutants. *Plant Cell* **9**, 409–423 (1997).
- Gosti, F. *et al.* *ABI1* protein phosphatase 2C is a negative regulator of absciscic acid signaling. *Plant Cell* **11**, 1897–1910 (1999).
- Li, J., Want, X., Watson, M. B. & Assmann, S. M. Regulation of absciscic acid-induced stomatal closure and anion channels by guard cell AAPK kinase. *Science* **287**, 300–303 (2000).
- Leung, J. & Giraudat, J. Absciscic acid signal transduction. *Ann. Rev. Plant Physiol. Plant Mol. Biol.* **49**, 199–222 (1998).
- Knetch, M. L. W., Wang, M., Snaar-Jagalska, B. E. & Heimovaara-Dijkstra, S. Absciscic acid induces mitogen-activated protein kinase activation in barley aleurone protoplasts. *Plant Cell* **8**, 1061–1067 (1996).
- Sheen, J. Ca²⁺-dependent protein kinases and stress signal transduction in plants. *Science* **274**, 1900–1902 (1996).
- Pawson, T. & Scott, J. D. Signaling through scaffold, anchoring, and adaptor proteins. *Science* **278**, 2075–2080 (1997).
- Ferl, R. J. 14-3-3 proteins and signal transduction. *Ann. Rev. Plant Physiol. Plant Mol. Biol.* **47**, 49–73 (1996).
- Kinoshita, T. & Shimazaki, K. Blue light activates the plasma membrane H⁺-ATPase by phosphorylation of the C-terminus in stomatal guard cells. *EMBO J.* **18**, 5548–5558 (1999).
- Leonhardt, N., Vasseau, A. & Forestier, C. ATP-binding cassette modulators control absciscic acid-regulated slow anion channels in guard cells. *Plant Cell* **11**, 1141–1151 (1999).
- Leonhardt, N., Marin, E., Vasseau, A. & Forestier, C. Evidence for the existence of a sulfonylurea-receptor-like protein in plants: Modulation of stomatal movements and guard cell potassium channels by sulfonylureas and potassium channel openers. *Proc. Natl Acad. Sci. USA* **94**, 14156–14161 (1997).
- Cutler, S., Ghassemian, M., Bonetta, D., Cooney, S. & McCourt, P. A protein farnesyl transferase involved in absciscic acid signal transduction in *Arabidopsis*. *Science* **273**, 1239–1241 (1996).
- Pei, Z. -M., Ghassemian, M., Kwak, C. M., McCourt, P. & Schroeder, J. I. Role of farnesyltransferase in ABA regulation of guard cell anion channels and plant water loss. *Science* **282**, 287–290 (1998).
- Yates, J. R. I. Mass spectrometry from genomics to proteomics. *Trends Genet.* **16**, 5–8 (2000).
- Plesch, G., Kamann, E. & Müller-Röber, B. Cloning of regulatory sequences mediating guard-cell-specific gene expression. *Gene* **249**, 83–89 (2000).
- Grabov, A. & Blatt, M. R. Co-ordination of signalling elements in guard cell ion channel control. *J. Exp. Bot.* **49**, 351–360 (1998).
- Allen, G. J. & Sanders, D. in *Advances in Botanical Research Incorporating Advances in Plant Pathology: The Plant Vacuole* (eds Leigh, R. A. & Sanders, D.) 217–252 (Academic Press, San Diego, 1997).
- Assmann, S. M. & Shimazaki, K. The multisensory guard cell. Stomatal responses to blue light and absciscic acid. *Plant Physiol.* **119**, 809–815 (1999).
- Pottosin, I. I., Tikhonova, L. I., Hedrich, R. & Schonknecht, G. Slowly activating vacuolar channels can not mediate Ca²⁺-induced Ca²⁺ release. *Plant J.* **12**, 1387–1398 (1997).
- Bewell, M. A., Maathuis, F. J. M., Allen, G. J. & Sanders, D. Calcium-induced calcium release mediated by a voltage-activated cation channel in vacuolar vesicles from red beet. *FEBS Lett.* **458**, 41–44 (1999).
- Wu, Y. *et al.* Absciscic acid signaling through cyclic ADP-ribose in plants. *Science* **278**, 2126–2130 (1997).

Acknowledgements

We thank Z.-M. Pei for discussions and help with figures and M. Ghassemian, N. Crawford, N. Spitzer, D. Waner and N. Leonhardt for comments. Supported by NSF, NIH and DOE grants (J.I.S.) and HFSP fellowships (J.M.K. and G.A.).

Structure of the bacterial flagellar protofilament and implications for a switch for supercoiling

Fadel A. Samatey*, Katsumi Imada*, Shigehiro Nagashima*, Ferenc Vonderviszt†, Takashi Kumasaka‡, Masaki Yamamoto‡ & Keiichi Namba*§

* Protonic NanoMachine Project, ERATO, JST, 3-4 Hikaridai, Seika, Kyoto 619-0237, Japan

† Department of Physics, University of Veszprém, Egyetem, u.10 H-8201, Hungary

‡ RIKEN Harima Institute, 1-1-1 Kouto, Mikazuki, Hyogo 679-5198, Japan

§ Advanced Technology Research Laboratories, Matsushita Electric Industrial Co., Ltd, 3-4 Hikaridai, Seika, Kyoto 619-0237, Japan

The bacterial flagellar filament is a helical propeller constructed from 11 protofilaments of a single protein, flagellin. The filament switches between left- and right-handed supercoiled forms when bacteria switch their swimming mode between running and tumbling. Supercoiling is produced by two different packing interactions of flagellin called L and R. In switching from L to R, the intersubunit distance (~52 Å) along the protofilament decreases by 0.8 Å. Changes in the number of L and R protofilaments govern supercoiling of the filament. Here we report the 2.0 Å resolution crystal structure of a *Salmonella* flagellin fragment of relative molecular mass 41,300. The crystal contains pairs of antiparallel straight protofilaments with the R-type repeat. By simulated extension of the protofilament model, we have identified possible switch regions responsible for the bi-stable mechanical switch that generates the 0.8 Å difference in repeat distance.

Bacteria swim by rotating helical flagellar filaments, which are up to 15 µm long, but only 120–250 Å in diameter. The rotary motor at the base of the filament drives the rotation of this helical propeller^{1,2} at hundreds of revolutions per second^{3,4}. For chemotaxis and thermotaxis, the swimming pattern of bacteria such as *Salmonella* and *Escherichia coli* alternates between 'run' and 'tumble'; a run lasts for a few seconds and a tumble for a fraction of second. During a run, the motor rotates counterclockwise (as it is viewed from outside the cell), and several flagellar filaments with a left-handed helical shape form a bundle and propel the cell. A tumble is caused by quick reversal of the motor to clockwise rotation⁵, which produces a twisting force that transforms the left-handed helical form of the filament into a right-handed one^{6,7}, causing the bundle to fall apart rapidly. The separated filaments act in an uncoordinated way to generate forces that change the orientation of the cell. Thus, the structure of the flagellar filament and its dynamic properties have an essential role in bacterial taxis.

The filament is a helical assembly of a single protein, flagellin, with roughly 11 subunits per 2 turns of the 1-start helix; the filament can also be described as a tubular structure comprising 11 strands of protofilaments, which are nearly longitudinal helical arrays of subunits⁸. The helical forms of the filament are caused by supercoiling, which is proposed to occur through a switching of conformation or packing interactions of the subunits between two distinct states and their non- and quasi-equivalent intersubunit interactions^{9–12}. In the proposed models, each protofilament exists in one of two slightly different conformations, which affect the repeat distance and lateral packing interaction; the regulated switching of the 11 protofilaments would then produce ten types of supercoil and two types of straight filament.

Many of these forms have actually been observed in various strains and under various conditions^{13–15}. The two straight filaments are called L-type and R-type according to the left- and right-handed twist of the protofilaments, respectively¹⁶. Accurate measurements by X-ray fibre diffraction show that the repeat distances of the L- and R-type protofilaments are 52.7 Å and 51.9 Å, respectively, the difference being only 0.8 Å. A simple mechanical simulation of

supercoiled forms with these structural parameters shows a good agreement with the pitch and diameter of observed supercoils, indicating that the data and the model are both physically sound¹⁷. This also indicates that the level of precision at which flagellin works as a lengthwise mechanical switch is on the order of a tenth of an ångström.

Electron cryomicroscopy and X-ray fibre diffraction have revealed the domain organization of flagellin and subunit packing in the two types of straight filaments at about 10 Å resolution^{17–22}; however, higher resolution is needed to understand the structural basis of supercoiling in atomic detail. Here we present the crystal structure of the flagellin F41 fragment (relative molecular mass (M_r) 41,300 (41K)) at 2.0 Å resolution, which directly reveals the protofilament structure. A portion of the F41 molecule involved in the axial subunit packing is associated with conformational switching, providing insights into a mechanical switch that works with sub-ångström precision.

Structure of F41

Because flagellin polymerizes into filaments, evading crystallization efforts, we prepared and crystallized a 41K fragment of *Salmonella* flagellin by clipping off peptides from both the amino- and carboxy-terminal ends. Although the crystals were only several micrometres thick, various improvements in the cryocrystallographic technique²³ allowed data collection at ESRF and SPring-8, resulting in a refined atomic model at 2.0 Å resolution.

The Cα backbone trace of F41 is shown in Fig. 1a. The overall shape of the molecule looks like a boomerang or an aircraft with two wings and a short body, each wing being about 70 Å long, 25 Å wide and 20 Å thick. The F41 structure can be divided into three domains, labelled D1, D2 and D3. Domain D1 comprises an N-terminal segment from Asn 56 to Gln 176, and a C-terminal segment from Thr 402 to Arg 450. Domain D2 also comprises two segments: Lys 177 to Gly 189, and Ala 284 to Ala 401. A central segment from Tyr 190 to Val 283 makes up domain D3. The domains are connected by short stretches of two chains in both cases. A cross β-motif is used to tie up the two ends of domain D1 connecting to

D2, where two hydrogen bonds are formed between Asn 173 and Thr 404. The two chains connecting domains D2 and D3, Gly 189 to Thr 193 and Val 281 to Asn 285, form a short β -strand. The distribution of domains and secondary structures along the amino-acid sequence is summarized in Fig. 1c.

Domain D1 is rod shaped, about 70 Å long and 20 Å wide, and comprises three α -helices and a strand of a unique β -turn/ β -hairpin/ α -helix motif. The N-terminal chain forms two α -helices and a β -hairpin. The first α -helix (residues 57–99) extends through the long axis of this domain, and the second one is relatively short (104–129). The β -hairpin (140–160), with a small crossed loop at its tip (146–153), is flanked by two distorted β -turns (131–134 and 135–138) on the N-terminal side, and slightly more than one turn of α -helix (163–168) on the C-terminal side. This α -helix is followed by an extended chain (170–176). The C-terminal chain forms an α -helix (406–447) as long as the first N-terminal α -helix. This rod-shaped domain forms an extensive hydrophobic core along its central axis in a similar way to four-helix bundles in other protein structures. But, because of an irregularity in one of the four strands involving the β -turn/ β -hairpin/ α -helix in series, the packing of hydrophobic side chains is relatively loose in the middle portion (Fig. 1b).

Domain D2 is made up of mostly β -strands except for one short 3_{10} helix (285–289) and one short α -helix (288–298). The fold of a set of these β -strands is unique and best described as a series of randomly oriented β -hairpins, with some of them also involved in three-stranded β -sheets. This domain has three distinct hydrophobic regions, two of which form enclosed hydrophobic cores; therefore, domain D2 can be divided into two compact subdomains, D2a and D2b (Fig. 1b).

Subdomain D2a consists of the N-terminal stretch (177–189) and the first half of the C-terminal segment (284–344) of domain

D2. Its hydrophobic core comprising 12 side chains is enclosed in a slightly irregular and unusual barrel formed by four β -strands and one α -helix. Subdomain D2b is formed by the last half of the C-terminal segment (345–401) of domain D2. Its hydrophobic core is relatively small, consisting of six side chains that are enclosed within the palm of a right-hand glove made of two β -hairpins corresponding to the thumb and forefinger and a loop corresponding to the remaining three fingers.

These two subdomains form domain D2 by intimate interactions between their hydrophobic surfaces. The third hydrophobic region of domain D2 is formed as part of this interface. These hydrophobic interactions between the two subdomains might explain why a proteolytic fragment composed of domains D2 and D3 behaves as two cooperatively unfolding domains instead of three in calorimetric analyses^{24,25}.

Domain D3 also consists of mostly β -strands with one short stretch of a helical fold (199–209); however, this helical stretch does not have the regular hydrogen-bonding pattern of an α -helix or a 3_{10} -helix. Domain D3 can be regarded as an unusual barrel made of four β -strands and one helical strand, which encloses 15 hydrophobic side chains in its core (Fig. 1a, b). This fold is similar to a part of D2a described above, and in addition a section of this fold can be described as a series of randomly oriented β -hairpins. Flagellin can have domain D3 deleted without losing its ability to form filaments, although the filaments are significantly more unstable than those of wild-type flagellin²⁶. The relatively independent arrangement of this domain from the rest explains why its deletion has a minimal effect on the structural integrity of flagellin. The reduction in the filament stability is due to the loss of intersubunit interactions, as shown by an electron microscopic study²².

In domains D2 and D3, the $C\alpha$ backbone trace shows a previously undescribed fold. Three portions with this fold are isolated and

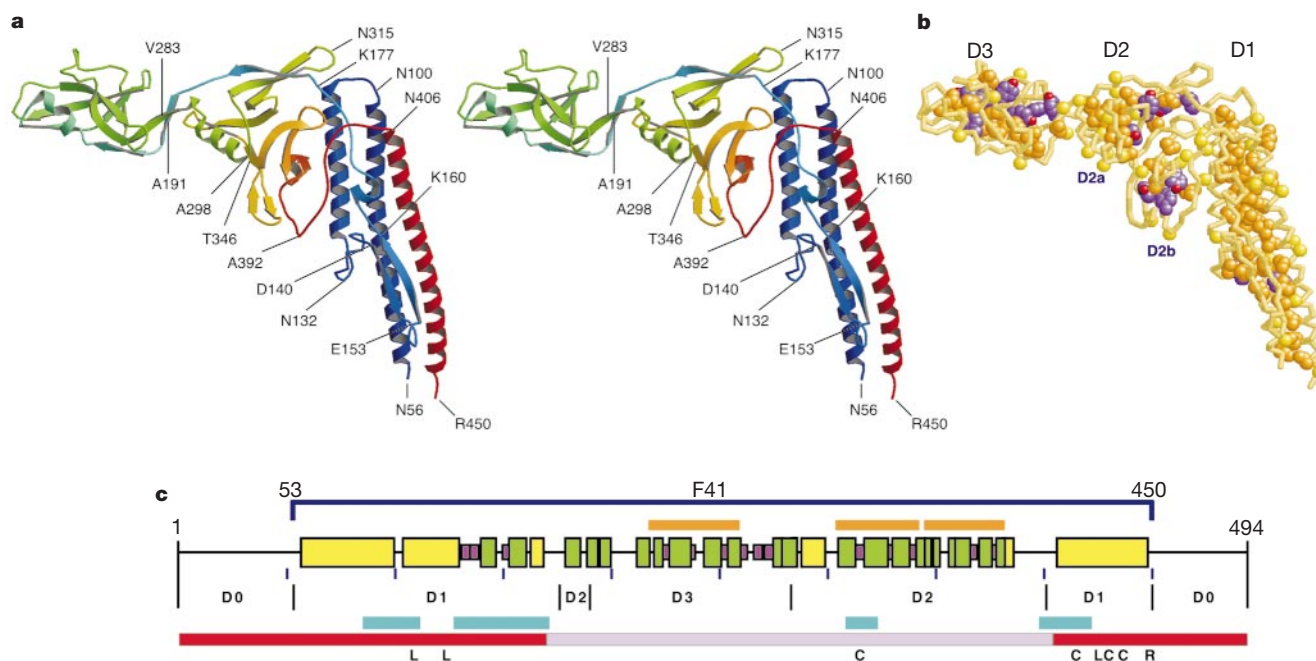


Figure 1 The $C\alpha$ backbone trace, hydrophobic core distribution and structural information of F41. **a**, Stereo diagram of the $C\alpha$ backbone. The chain is colour coded from blue to red from the N to the C terminus. Prepared with MOLSCRIPT⁴⁴ and RASTER3D⁴⁵. **b**, Four distinct hydrophobic cores that define domains D1, D2a, D2b and D3. All the hydrophobic side-chain atoms are displayed with the $C\alpha$ backbone. Side-chain atoms are colour coded: Ala, yellow; Leu, Ile or Val, orange; Phe and Tyr, purple (carbon atoms) and red (oxygen atoms). Prepared with RASMOL⁴⁶. **c**, Position and region of various structural features in the amino-acid sequence of flagellin. Shown are, from top to bottom: the F41

fragment in blue; three β -folium folds in brown; the secondary structure distribution with α -helix in yellow, β -structure in green, and β -turn in purple; tic mark at every 50th residue in blue; domains D0, D1, D2 and D3; the axial subunit contact region within the protofilament in cyan; the well-conserved amino-acid sequence in red and variable region in violet; point mutations in F41 that produce the filaments of different supercoils. Letters at the bottom indicate the morphology of mutant filaments: L (D107E, R124A, R124S, G426A), L-type straight; R (A449V), R-type straight; C (D313Y, A414V, A427V, N433D), curly³³.

shown in Fig. 2 (corresponding regions of the amino-acid sequence are indicated in Fig. 1c). The fold comprises a series of β -hairpins all pointing away from one another like three thin leaves spread out or like a curve of cubic polynomial called the 'folium of Descartes'. We therefore named it the β -folium. A common feature found in these β -foliums is that most of the tips of β -hairpins are either bent or twisted, and so some of them are better described as a β -finger with a claw.

Structure of the protofilament

The flagellar filament is made of 11 protofilaments all with the same polarity. From fibre-diffraction studies, two distinct subunit repeat distances along the protofilament have been identified as L- and R-type¹⁶. More accurate measurements have determined that the L-type repeat is 52.7 Å and the R-type repeat is 51.9 Å, with experimental errors smaller than 0.1 Å (ref. 17). The repeat distance along the *a* axis of the F41 crystal is 51.8 Å (± 0.1 Å), which encouraged us to look for the protofilament structure in the crystal as soon as the electron density map became available. As expected, we found that the protofilament structure with F41 molecules lined up along the *a* axis in the same orientation. The overall shape of the protofilament looks exactly the same as those found in lower resolution density maps obtained by electron cryomicroscopy^{18,19}.

The molecular packing in the crystal is shown in Fig. 3 in two orthogonal views. Because the crystal has a space group of $P2_1$, two perfectly straight protofilaments are packed antiparallel and this pair repeats along the *b* axis every 36.5 Å, forming a flat sheet. This mode of molecular packing is the same as that found in the zinc sheet of tubulin, in which the protofilaments of microtubule form an antiparallel array within the monolayer sheet²⁷. The flagellar protofilaments formed a stack of sheets up to a few hundred layers with a repeat of 119 Å, producing three-dimensional, plate crystals that were several micrometres thick.

The atomic model of one protofilament from the F41 crystal structure was relatively easily docked onto a low-resolution density map of the R-type straight filament from electron cryomicroscopy and fibre diffraction^{17,18} (Fig. 4). The fit is nearly perfect for a few consecutive subunits, clearly showing that the protofilament structure in the crystal is almost identical to that in the filament. Although the protofilament model is straight and the corresponding protofilament density is gently wound into a right-handed helix, the twist and curvature of this helix are very small.

Three of the four domains identified previously in the filament density map, D1, D2 and D3 (named in the order of their radial positions from inside to outside), correspond with the three domains identified in the F41 structure. Domain D1 of the F41 atomic model is located in the outer-tube region of the filament (Fig. 4, OT), filling most of the outer-tube volume, but not the inner-tube region (Fig. 4, IT). This indicates that a large portion of the truncated terminal regions from Ala 1 to Arg 52 and from Ser 451 to Arg 494, which are unfolded in the monomeric form, occupies the inner-tube volume. Domains D2 and D3 respectively

fill the two outer domains of the density map. These observations are consistent with previous amino-acid assignments of the domains^{17,18,20–22,28}, with slight corrections to the domain boundaries (Fig. 1c). The nearly axial alignment of long α -helices in the outer-tube region matches the original predictions from X-ray fibre diffraction data²⁹.

The outer-tube domain is known to be solely responsible for the formation of two distinct helical lattices, L-type and R-type²⁰. Even when the inner-tube structures are removed by truncating the terminal segments, flagellin fragments can polymerize into supercoiled filaments if seeded by the wild-type supercoiled filaments³⁰. This means that the outer-tube domain D1 has the ability to switch between the two states of flagellin represented by the L-type and R-type straight filament structures.

The protofilament of the R-type repeat is held in the crystal by axial interactions between D1 domains, and it exists in the absence of lateral protofilament interactions that construct the flagellar filament. This indicates that each protofilament is an independent, cooperatively switching unit, at least for the lengthwise mechanical switch that produces the curvatures of supercoiled filaments by incorporating the two types of protofilament into the tubular structure. This also indicates that the axial intersubunit interactions

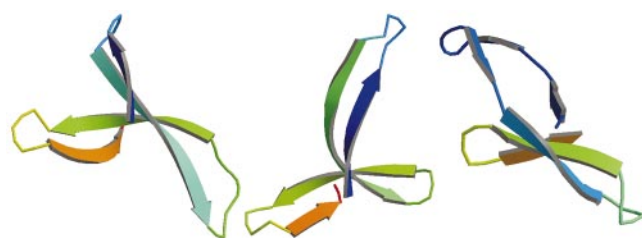


Figure 2 The β -folium fold. Left, residues 220–260 in domain D3; middle, residues 308–345 in domain D2a; right, residues 345–383 in domain D2b. The chain is colour coded from blue to green to brown according to the amino-acid sequence. Prepared with MOLSCRIPT⁴⁴ and RASTER3D⁴⁵.

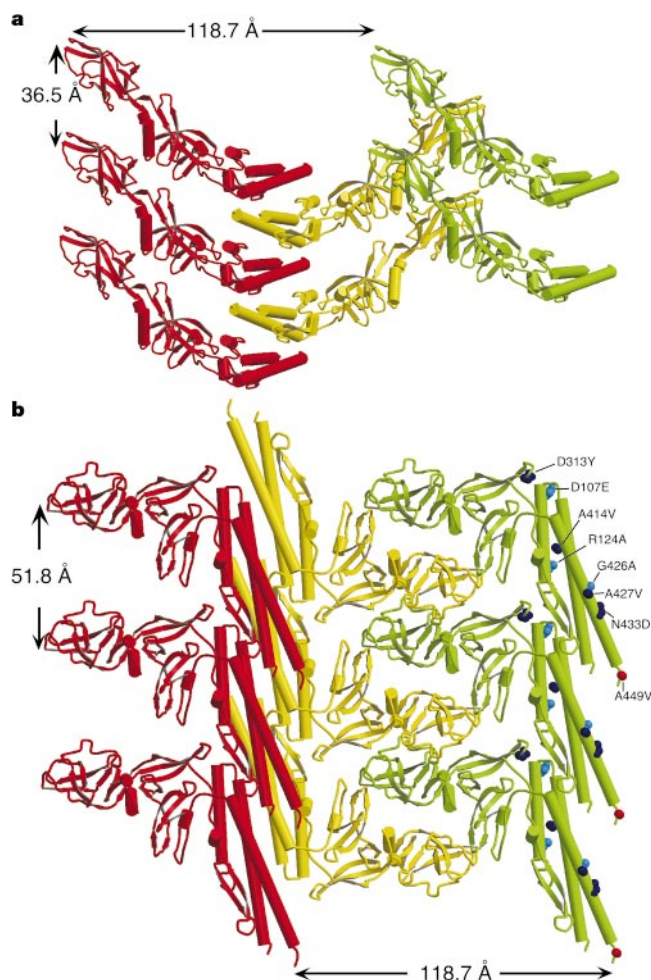


Figure 3 Crystal packing of F41 in two orthogonal views. **a**, The *b*–*c* plane viewed down the *a* axis. **b**, The *a*–*c* plane viewed down the *b* axis. Each array of the F41 molecules along the *a* axis, which has been identified as the protofilament of the flagellar filament, is coloured red, yellow and green. The repeat distances are labelled. Antiparallel packing of the protofilaments by the $P2_1$ symmetry of the crystal is clearly shown. Mutation sites that affect the supercoiling are also labelled with side-chain atoms shown in CPK representation. Mutations are coloured blue for L-type, red for R-type, and black for curly. Prepared with MOLSCRIPT⁴⁴ and RASTER3D⁴⁵.

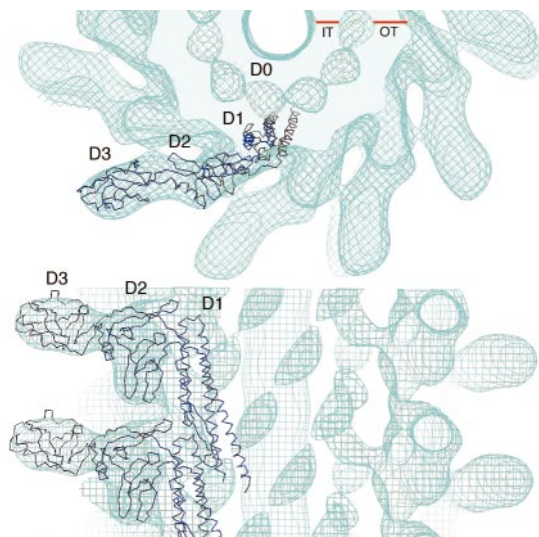


Figure 4 Docking of a single protofilament model into the density map of the filament. Top panel, end-on view from the top, with a 50 Å thick cross-section of the map. The map is painted where the density is higher than the contour level. Bottom panel, side view, with a 50 Å thick longitudinal section of the map that spans from the filament axis towards readers. The top points towards the distal end of the flagellum. The resolution of the density map was limited to 20 Å for easy evaluation of the model fitting. D0, D1, D2 and D3 indicate the domains of flagellin; IT and OT represent the inner- and outer-tube regions, respectively. Prepared with O⁴¹.

between D1 domains along the protofilament are responsible for the two-state switching to produce the two distinct repeat distances.

Axial interactions within the protofilament

The axial intersubunit interactions are formed between domain D1 of the upper subunit and domain D1 and a small portion of D2a of

the lower subunit (Fig. 5a). The contact surface is relatively small, and that is probably why, unlike the case of disassembling microtubule tips³¹, isolated protofilaments have never been observed under physiological conditions. The protofilament seems to have sufficient stability only in the presence of lateral interactions: parallel in the filament structure and antiparallel in the crystal. A magnified image of the axial interface is shown in Fig. 5b. In the bottom part of the upper subunit, a short segment of the first N-terminal α -helix (Asn 56 to Asp 69) and the unique motif of two consecutive β -turns/ β -hairpin (Phe 132 to Asp 151) form a concave surface. In the top part of the lower subunit, a convex surface is formed by short segments of two N-terminal α -helices with a short loop connecting these α -helices (Gln 89 to Asp 107), Leu 408 and Gln 409 of the C-terminal α -helix, and Asn 315 of subdomain D2a. These two surfaces have complementary shapes, which produce a number of van der Waals contacts.

The nature of the interactions is mainly polar–polar or charge–polar. There are two small patches where the interactions are formed between polar and hydrophobic groups, but there are no hydrophobic–hydrophobic interactions. Charge–polar interactions are formed between Asp 69 and Asn 100, Asn 132 and Asp 107, the carbonyl oxygen of Ala 149 and Arg 92, and Asp 151 and Gln 409, of the upper and lower subunit, respectively. Thus, two of the four residues, Ala 149 and Asp 151, are in the short loop at the tip of the β -hairpin of the upper subunit, where Asp 151 is capping the N-terminal end of the C-terminal α -helix of the lower subunit. A portion of the β -hairpin also forms a short intersubunit β -strand with a segment just after the first N-terminal α -helix of the lower subunit. These indicate the important roles of the β -hairpin in domain D1 (Asp 140 to Lys 160) in these axial interactions.

Mechanical switching unit

As discussed above, the protofilament itself is likely to have the ability to switch between the two mechanically stable states with the L- and R-type repeat. As the L-type repeat is only 0.8 Å longer than the R-type, the conformational change would be relatively small. We

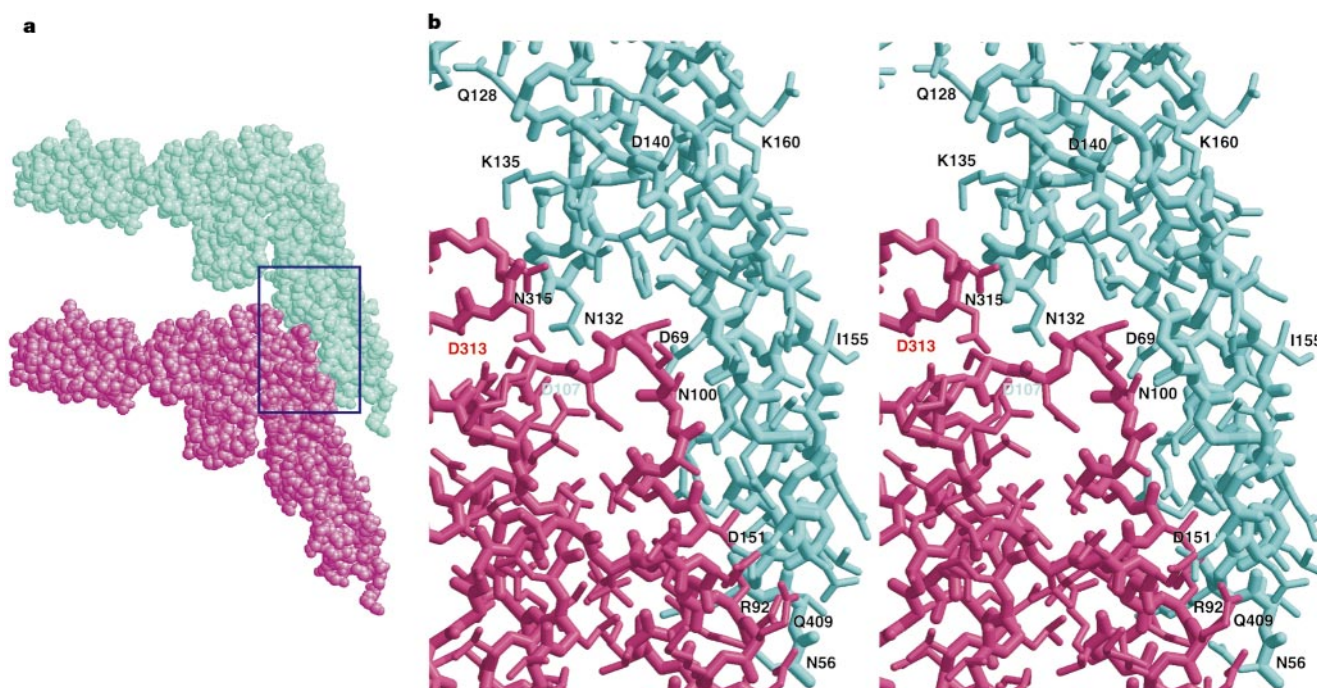


Figure 5 Axial interactions in the protofilament. **a**, Space-filling representation showing the axial packing of two subunits. Prepared with RASMOL⁴⁶. **b**, Stereo close-up view of the boxed region in **a**. The two subunits shown are coloured cyan (upper) and dark pink

(lower). Some of the residues are labelled to guide identification. Prepared with MOLSCRIPT⁴⁴ and RASTER3D⁴⁵.

looked for a local conformational switch in the atomic model of F41 by simulating a gradual extension of the protofilament by fixing its one end and pulling the other end in the axial direction. We used a three-subunit protofilament model, and fixed the top subunit while pulling down the bottom subunit to see what happened in the middle subunit. At every step we translated the bottom subunit by 0.1 Å with its C α backbone as a rigid body and carried out energy minimization of the model.

The results are shown in Fig. 6. Up to 4.5 Å, every portion of the middle subunit was elastically stretched gradually, as typically shown by the elongating pitch of α -helices (Fig. 6a). But immediately after this point, further translation over the next two steps (0.2 Å) caused an abrupt change in the conformation of the β -hairpin in domain D1, which was accompanied by a small movement of the C-terminal α -helix. Further translation caused further gradual distortion of the molecule as observed before the transition. The two conformations before and after the transition are superimposed for direct comparison in Fig. 6b. The characteristic nature of this switching is a small but significant movement of the β -hairpin that slightly pushes out and down the lower subunit at the axial interface, making the axial subunit dispositions with a longer repeat distance stable.

The two terminal α -helical segments more or less follow the β -hairpin movement, which probably maintains the side-chain interactions with the β -hairpin in the hydrophobic core. Because we carried out the simulated extension with an isolated three-subunit protofilament, without having all those restraints from lateral protofilament packing that are supposed to be present in the filament structure, some of the conformational changes shown here may be artefacts. A more thorough model simulation is

needed to prove that the β -hairpin is the switch; however, our current result strongly suggests that the conformational switching of the β -hairpin is responsible for the two distinct states of flagellin packing with two distinct repeat distances at sub-ångström accuracy.

Mutations that affect the supercoiling

The amino-acid sequences of flagellin from various bacterial strains show two extremely well conserved regions, which are about 170 residues from the N terminus and about 90 residues from the C terminus (for example, by ProDom domain comparison³²). From the F41 structure and its position in the filament density, it is clear why these regions are highly conserved. These conserved regions form the densely packed core of the filament—the outer and inner tube (Fig. 4). About a dozen point mutation sites have been identified for a wild-type strain of *S. typhimurium* SJW1103 in relation to the filament morphology³³, and all except one are found within these conserved regions (Fig. 1c). Within F41, mutations D107E, R124A, R124S, G426A and A449V give rise to straight filaments; and mutations D313Y, A414V, A427V, N433D and A449T result in curly filaments (see Fig. 3). We have attempted to interpret the effect of these mutations on the basis of the protofilament structure.

Asp 107 is located at the beginning of the second N-terminal α -helix in the top portion of the lower subunit, and interacts with Asn 132 in the first of the two consecutive β -turns before the β -hairpin in domain D1 of the upper subunit (Fig. 5). The mutant flagellin D107E (from strain SJW1663) forms the L-type straight filament³⁴. In the protofilament model, the side chain of Asp 107 points up whereas that of Asn 132 points down, and the distance

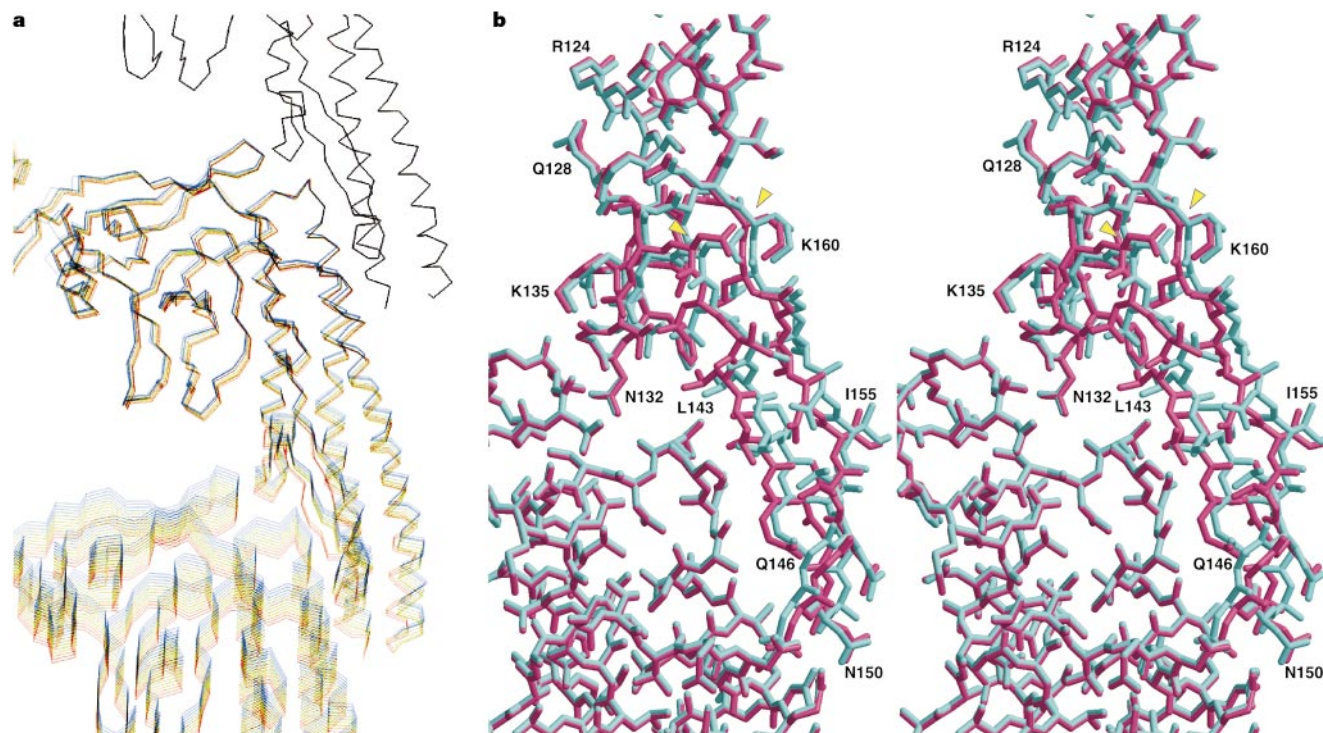


Figure 6 Simulated extension of the protofilament showing the possible switch region. **a**, Superimposition of 12 sampled stages of the simulated extension. Domains D1 and D2 of three F41 molecules are shown (bottom half of the top, whole of the middle, and top half of the bottom) and the 12 stages are colour coded from blue to red. Each step of the extension was 0.1 Å, but every five steps is sampled and displayed here; therefore, the bottom subunits in different stages are equally separated by 0.5 Å. The whole β -hairpin at the bottom of domain D1 of the middle subunit shows a discrete jump in its conformation

at an extension from 4.5 to 5.0 Å, whereas other parts show only gradual deformation of an elastic nature. Prepared with O⁴¹. **b**, Close-up view of the axial contact region in stereo, including the β -hairpin showing the conformational switch. The structures at two sampled stages, at an extension of 4.5 Å in cyan and of 4.7 Å in dark pink, are superimposed to show the conformational switch of the β -hairpin, of which the start and the end are marked by yellow arrowheads. Prepared with MOLSCRIPT⁴⁴ and RASTER3D⁴⁵.

Table 1 Summary of refinement statistics

Resolution range	No. of reflections with $F > 2\sigma$	No. of residues	No. of protein atoms	No. of water molecules	R_{cryst}^*	$R_{\text{free}}^\dagger$	r.m.s. bond length	r.m.s. bond angle
10–2.0 Å (2.09–2.0)	29,866 (3,615)	395	2,880	354	23.4% (30.9%)	26.6% (30.8%)	0.009 Å	1.5°

* $R_{\text{cryst}} = \sum |F_{\text{obs}}(hkl) - F_{\text{calc}}(hkl)| / \sum F_{\text{obs}}(hkl)$.

† As R_{cryst} , but calculated on 8% of data set aside for refinement. Numbers in parentheses refer to those in the highest resolution shell.

between them is slightly longer than an average hydrogen bond. Mutation from aspartate to glutamate, which has a slightly longer side chain, would strengthen this hydrogen bond, which may stabilize the conformation of the switch region in the L-state.

Asp 313 is the only mutation located away from the conserved regions, and was previously puzzling. The involvement of this residue is now clearly shown in the structure. Asp 313 is close to Asn 315 in a β -hairpin of subdomain D2a of the lower subunit, where Asn 315 makes a hydrogen bond with the main-chain oxygen of Gly 133 of the upper subunit, which is also part of the β -turn to which Asn 132 belongs (Fig. 5). Although not directly, Asp 313 is involved in the axial packing interactions by three portions of the molecule: the β -turn of the upper subunit; the loop between the two N-terminal α -helices; and the tip of the β -hairpin of subdomain D2a of the lower subunit. The mutation D313Y gives rise to curly filaments—one of the right-handed supercoils produced by increasing the number of the R-type protofilaments from two to four or five (a normal filament of wild-type flagellin, a left-handed supercoil, consists of two R-type and nine L-type protofilaments). Because tyrosine has a significantly bulkier side chain than has aspartic acid, this may further stabilize the packing of the three portions in the R-type conformation.

Ala 414 is involved in the hydrophobic core of domain D1 and is close to the short α -helix after the β -hairpin. Its replacement with valine also results in curly filaments, probably through stabilization of the hydrophobic core, which in turn stabilizes the R-type conformation of the switch region.

The other mutated residues do not have any bonding partners within the F41 protofilament; Arg 124 and Ala 427 probably face the lateral neighbours, and Gly 426, Asn 433 and Ala 449 appear to point toward the inner-tube domain, which is missing in the current model. This indicates that the specific form of supercoiling is determined mostly by the lateral interactions between the protofilaments and/or the interactions between the outer- and inner-tube domains.

Conformation of flagellin during transport

For filament growth, flagellin molecules are transported to the distal end of the flagellum through its central channel, which has a diameter of about 30 Å (refs 18, 19). Although there is no data on the conformation of flagellin during the transport, flagellin obviously either has to be unfolded or at least have its domain largely rearranged.

The structure of F41 suggests that, even with intact terminal regions, flagellin is made of independent domains that are connected linearly, and that each domain except for domain D2 is thin enough to fit inside the channel. Therefore, it may be sufficient to have partially unfolded domain D2 and flexible interdomain connections during the transport. If flagellin comes out of the flagellum through the opening under the cap at the distal end³⁵ in the order of D3 first followed by D2 and D1, it would be ideal for a rapid self-assembly process. This question remains to be answered by experimental data.

Towards construction of the filament model

The structure of the flagellar protofilament has provided many insights into the structural mechanism of flagellin polymerization and the possible switch region responsible for the two-state

mechanical switch of sub-ångstrom precision, which defines the curvatures of the supercoiled flagellar filaments. However, understanding the mechanisms that determine a particular twist of a supercoil and its dynamic switching between left- and right-handed ones, for example by the twisting force produced by rapid motor reversal, requires detailed analyses of the lateral packing interactions between the protofilaments in the filament structure.

The approximate models of the L- and R-type straight filaments produced by docking the protofilament model into low-resolution density maps with slight deformations are not sufficient, because the switching in the lateral interactions is a mutual sliding over a few ångströms¹⁷. To identify correctly the atomic interactions involved in the switching, we are now incorporating experimental data obtained from electron cryomicroscopy and X-ray fibre diffraction with those obtained from X-ray crystallography to construct accurate models of the L- and R-type straight filaments. □

Methods

F41 preparation and crystallization

The F41 fragment of flagellin from *Salmonella typhimurium* was prepared and crystallized as described²³. Briefly, crystals were grown in solutions containing 6% PEG-6000, 12% glycerol, 3–6% isopropanol, 50–75 mM NaCl, 20 mM Tris-HCl, pH 7.8, at 16 °C. Glycerol and isopropanol were important in reducing the otherwise very high nucleation rate, which resulted in many micrometre-sized needle crystals²³.

Data collection

The space group of the crystal was $P2_1$, with the cell dimensions $a = 51.8$, $b = 36.5$, $c = 118.7$ Å, $\beta = 90.8^\circ$. The solvent content was about 55% and the crystal contained one molecule per asymmetric unit. The size of the F41 crystal was typically $0.5 \text{ mm} \times 0.2 \text{ mm} \times <0.01 \text{ mm}$, which made data collection difficult. The crystals were frozen in liquid propane and routinely annealed in the cryostream to improve the diffraction at high resolution²³. All the data sets were collected at temperatures around 100 K. Because of the long exposure time (5 h per frame) required to record high-resolution spots, even with the high-brilliance X-ray beam of our in-house Rigaku X-ray oscillation camera (fine focus X-ray generator, FR-D; Yale-type optics; image plate detector, RAXIS-IV), only well-diffracting crystals were selected in the laboratory, and full data sets were collected at synchrotron beam lines. Data sets used for the structure analysis are listed in the Supplementary Information.

A set of multi-wavelength anomalous diffraction (MAD) data, which produced a high-quality electron density map at 2.0 Å resolution, were collected under Trichromatic concept³⁶ at RIKEN beamline I (BL45XU) at the 8 GeV Super Photon ring (SPring-8) in Harima³⁷. The data were reduced by DENZO and SCALEPACK³⁸, or MOSFLM³⁹ and SCALA⁴⁰. The atomic model was built as shown in the Supplementary Information. The model was refined at 2.0 Å resolution including 354 water molecules (Table 1).

Structure analysis

Because the F41 crystals prepared from three different flagellins (wild type, L-type, and G365C mutant of the R-type) all showed the same cell dimensions within experimental errors, we treated them as isomorphous crystals. An electron density map was first obtained at 2.5 Å resolution by multiple isomorphous replacement using two derivative data sets (LS7 and C77; see Supplementary Information). We traced the main chain in domain D1 using the graphics program O⁴¹, but it was difficult to complete the model building. Then, we obtained new phases to 2.0 Å using a set of MAD data (G19; see Supplementary Information) and program SOLVE⁴², which enabled us to complete the model building. The N-terminal three residues are not included in the model because that part of electron density was obscure. We then refined the atomic model, comprising 395 amino-acid residues and 354 water molecules, against the data collected at ESRF ID14-3 by using X-PLOR⁴³ (see Table 1). We also refined the model against three data sets obtained from the F41 crystals prepared from three different flagellins, but could not find any significant difference between the three, which we will describe and discuss in more detail elsewhere.

Simulated extension of the protofilament

Simulation of the protofilament extension was carried at every 0.1 Å step by energy minimization of the three-subunit protofilament model with C α atoms of both top and

bottom subunits treated as rigid bodies. We used a routine in X-PLOR⁴³ for energy minimization.

Received 2 November 2000; accepted 11 January 2001.

1. Berg, H. C. & Anderson, R. A. Bacteria swim by rotating their flagellar filaments. *Nature* **245**, 380–382 (1973).
2. Silverman, M. & Simon, M. Flagellar rotation and the mechanism of bacterial motility. *Nature* **249**, 73–74 (1974).
3. Kudo, S., Magariyama, Y. & Aizawa, S.-I. Abrupt changes in flagellar rotation observed by laser dark-field microscopy. *Nature* **346**, 677–680 (1990).
4. Ryu, W. S., Berry, R. M. & Berg, H. C. Torque-generating units of the flagellar motor of *Escherichia coli* have a high duty ratio. *Nature* **403**, 444–447 (2000).
5. Larsen, S. H., Reader, R. W., Kort, E. N., Tso, W. W. & Adler, J. Change in direction of flagellar rotation is the basis of the chemotactic response in *Escherichia coli*. *Nature* **249**, 74–77 (1974).
6. Macnab, R. M. & Ornston, M. K. Normal-to-curl flagellar transitions and their role in bacterial tumbling. Stabilization of an alternative quaternary structure by mechanical force. *J. Mol. Biol.* **112**, 1–30 (1977).
7. Turner, L., Ryu, W. S. & Berg, H. C. Real-time imaging of fluorescent flagellar filaments. *J. Bacteriol.* **182**, 2793–2801 (2000).
8. O'Brien, E. J. & Bennett, P. M. Structure of straight flagella from a mutant *Salmonella*. *J. Mol. Biol.* **70**, 133–152 (1972).
9. Asakura, S. Polymerization of flagellin and polymorphism of flagella. *Adv. Biophys. (Japan)* **1**, 99–155 (1970).
10. Calladine, C. R. Construction of bacterial flagella. *Nature* **225**, 121–124 (1975).
11. Calladine, C. R. Design requirements for the construction of bacterial flagella. *J. Theor. Biol.* **57**, 469–489 (1976).
12. Calladine, C. R. Change of waveform in bacterial flagella: The role of mechanics at the molecular level. *J. Mol. Biol.* **118**, 457–479 (1978).
13. Kamiya, R. & Asakura, S. Helical transformations of *Salmonella* flagella in vitro. *J. Mol. Biol.* **106**, 167–186 (1976).
14. Kamiya, R. & Asakura, S. Flagellar transformations at alkaline pH. *J. Mol. Biol.* **108**, 513–518 (1977).
15. Hotani, H. Micro-video study of moving bacterial flagellar filaments III. Cyclic transformation induced by mechanical force. *J. Mol. Biol.* **156**, 791–806 (1982).
16. Kamiya, R., Asakura, S., Wakabayashi, K. & Namba, K. Transition of bacterial flagella from helical to straight forms with different subunit arrangements. *J. Mol. Biol.* **131**, 725–742 (1979).
17. Yamashita, I. *et al.* Structure and switching of bacterial flagellar filament studied by X-ray fiber diffraction. *Nature Struct. Biol.* **5**, 125–132 (1998).
18. Mimori, Y. *et al.* The structure of the R-type straight flagellar filament of *Salmonella* at 9 Å resolution by electron cryomicroscopy. *J. Mol. Biol.* **249**, 69–87 (1995).
19. Morgan, D. G., Owen, C., Melanson, L. A. & DeRosier, D. J. Structure of bacterial flagellar filaments at 11 Å resolution: Packing of the α -helices. *J. Mol. Biol.* **249**, 88–110 (1995).
20. Mimori-Kiyosue, Y., Vonderviszt, F., Yamashita, I., Fujiyoshi, Y. & Namba, K. Direct interaction of flagellin termini essential for polymorphic ability of flagellar filament. *Proc. Natl Acad. Sci. USA* **93**, 15108–15113 (1996).
21. Mimori-Kiyosue, Y., Vonderviszt, F. & Namba, K. Locations of terminal segments of flagellin in the filament structure and their roles in polymerization and polymorphism. *J. Mol. Biol.* **270**, 222–237 (1997).
22. Mimori-Kiyosue, Y., Yamashita, I., Fujiyoshi, Y., Yamaguchi, S. & Namba, K. Role of the outermost subdomain of *Salmonella* flagellin in the filament structure revealed by electron cryomicroscopy. *J. Mol. Biol.* **284**, 521–530 (1998).
23. Samatey, F. A., Imada, K., Vonderviszt, F., Shirakihara, Y. & Namba, K. Crystallization of the F41 fragment of flagellin and data collection from extremely thin crystals. *J. Struct. Biol.* **132**, 106–111 (2000).
24. Vonderviszt, F., Uedaira, H., Kidokoro, S.-I. & Namba, K. Structural organization of flagellin. *J. Mol. Biol.* **214**, 97–104 (1990).
25. Honda, S., Uedaira, H., Vonderviszt, F., Kidokoro, S.-I. & Namba, K. Folding energetics of a multidomain protein, flagellin. *J. Mol. Biol.* **293**, 719–732 (1999).
26. Yoshioka, K., Aizawa, S.-I. & Yamaguchi, S. Flagellar filament structure and cell motility of *Salmonella typhimurium* mutants lacking part of the outer domain of flagellin. *J. Bacteriol.* **177**, 1090–1093 (1995).
27. Nogales, E., Sharon, G. W., & Downing, K. H. Structure of the $\alpha\beta$ tubulin dimer by electron crystallography. *Nature* **391**, 199–203 (1998).
28. Yamashita, I. *et al.* Radial mass analysis of the flagellar filament of *Salmonella*: Implications for subunit folding. *J. Mol. Biol.* **253**, 547–558 (1995).
29. Namba, K., Yamashita, I. & Vonderviszt, F. Structure of the core and central channel of bacterial flagella. *Nature* **342**, 648–654 (1989).
30. Vonderviszt, F., Aizawa, S.-I. & Namba, K. Role of the disordered terminal regions of flagellin in filament formation and stability. *J. Mol. Biol.* **221**, 1461–1474 (1991).
31. Mandelkow, E. -M., Mandelkow, E. & Milligan, R. A. Microtubule dynamics and microtubule caps: A time-resolved cryo-electron microscopy study. *J. Cell Biol.* **114**, 977–991 (1991).
32. Corpet, F., Gouzy, J. & Kahn, D. The ProDom database of protein domain families. *Nucleic Acids Res.* **26**, 323–326 (1998).
33. Kanto, S., Okino, H., Aizawa, S.-I. & Yamaguchi, S. Amino acids responsible for flagellar shape are distributed in terminal regions of flagellin. *J. Mol. Biol.* **219**, 471–480 (1991).
34. Kamiya, R., Asakura, S. & Yamaguchi, S. Formation of helical filaments by copolymerization of two types of 'straight' flagellins. *Nature* **286**, 628–630 (1980).
35. Yonekura, K. *et al.* The bacterial flagellar cap as the rotary promoter of flagellin self-assembly. *Science* **290**, 2148–2152 (2000).
36. Yamamoto, M., Kumasaka, T., Fujisawa, T. & Ueki, T. Trichromatic concept at Spring-8 RIKEN beamline I. *J. Synchrotron Rad.* **5**, 222–225 (1998).
37. Ueki, T. & Yamamoto, M. The start of a new generation: the present status of the Spring-8 synchrotron and its use in structural biology. *Structure* **7**, R183–R187 (1999).
38. Otwinowski, Z. & Minor, W. *Processing of X-ray Diffraction Data Collected in Oscillation Mode* (Academic, New York, 1997).
39. Leslie, A. G. W. CCP4/ESF-EACMB. *Newslett. Protein Crystallogr.* Vol. 26 (Daresbury Laboratory, Warrington, UK, 1992).
40. Collaborative Computational Project Number 4. The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
41. Jones, T. A., Zhou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
42. Terwilliger, T. C. & Berendzen, J. Automated structure solution for MIR and MAD. *Acta Crystallogr. D* **55**, 849–861 (1999).
43. Brünger, A. T., Kuriyan, J. & Karplus, M. Crystallography R factor refinement by molecular dynamics. *Science* **235**, 458–460 (1987).
44. Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
45. Merritt, E. A. & Bacon, D. J. Raster3D: Photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524. (1997).
46. Sayle, R. A. & Milner-White, E. J. RasMol: Biomolecular graphics for all. *Trends Biochem. Sci.* **20**, 374–376 (1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank T. Tomizaki, L. Dumon, W. Burmeister, S. Arzt and S. Wakatsuki at ESRF, and M. Kawamoto, N. Kamiya and K. Miura at SPring-8 for technical help with beamlines. We also thank I. Yamashita and K. Hasegawa for a mutant strain of *Salmonella* that produces SJW1655-derived site-directed mutant flagellin (G365C), which forms the R-type straight flagellar filament, and helpful information of heavy-atom binding to the filament; J. Tame for critically reading the manuscript; and S. Asakura, T. Nitta and F. Oosawa for continuous support and encouragement.

Correspondence and requests for materials should be addressed to K.N. (e-mail: keiichi@crl.mei.co.jp). Atomic coordinates have been deposited in the Protein Data Bank under accession code 1IO1.

Discovery of radio emission from the brown dwarf LP944–20

E. Berger*, S. Ball†, K. M. Becker‡, M. Clarke§, D. A. Frai||, T. A. Fukuda¶, I. M. Hoffman#, R. Mellon*, E. Momjian**, N. W. Murphy††, S. H. Teng‡‡, T. Woodruff§§, B. A. Zauderer||| & R. T. Zavala¶¶

* Division of Physics, Mathematics & Astronomy 105-24, California Institute of Technology, Pasadena, California 91125, USA

† Department of Physics, New Mexico Tech, PO Box 3394 c/s, Socorro, New Mexico 87801, USA

‡ Department of Physics, Oberlin College, Oberlin, Ohio 44074, USA

§ Department of Physics, Carleton College, 300 N. College St., Northfield, Minnesota 55057, USA

|| National Radio Astronomy Observatory, PO Box O, Socorro, New Mexico 87801, USA

¶ Department of Physics & Astronomy, University of Denver, 2112 E. Wesley Ave., Denver, Colorado 80208, USA

Department of Physics & Astronomy, University of New Mexico, 800 Yale Blvd NE, Albuquerque, New Mexico 87131, USA

** Department of Astronomy, 525 Davey Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

*** Department of Physics & Astronomy, 177 Chem-Physics building, University of Kentucky, Lexington, Kentucky 40506, USA

†† Department of Physics, Amherst College, Amherst, Massachusetts 01002-5000, USA

‡‡ Department of Astronomy, University of Maryland, College Park, Maryland 20742, USA

§§ Department of Physics, Southwestern University, SU Box 7263, Georgetown, Texas 78626, USA

||| Department of Physics & Astronomy, Agnes Scott College, Decatur, Georgia 30030, USA

¶¶ Department of Astronomy, New Mexico State University, MSC 4500 PO Box 3001, Las Cruces, New Mexico 88003, USA

Brown dwarfs are not massive enough to sustain thermonuclear fusion of hydrogen at their centres, but are distinguished from gas-giant planets by their ability to burn deuterium¹. Brown dwarfs older than ~10 Myr are expected to possess short-lived magnetic fields² and to emit radio and X-rays only very weakly from their coronae. An X-ray flare was recently detected³ on the brown dwarf LP944–20, whereas previous searches^{4–7} for optical activity (and one X-ray search¹) yielded negative results. Here we report the discovery of quiescent and flaring radio emission from LP944–20, with luminosities several orders of magnitude larger than predicted by the empirical relation^{8,9} between the X-ray and radio luminosities that has been found for many types of stars. Interpreting the radio data within the context of synchrotron emission, we show that LP944–20 has an unusually weak magnetic field in comparison to active M-dwarf stars^{10,11}, which might explain the previous null optical^{4–7} and X-ray¹ results, as well as the strength of the radio emissions compared to those at X-ray wavelengths.

Following the detection of an X-ray flare from the brown dwarf LP944–20 on 15 December 1999 by the Chandra X-ray Observatory³, we observed the X-ray position of this source with the Very Large Array (VLA) on 2000 July 27.54 UT, and detected a radio source. We tested the radio emission for variability by imaging subsets of the observation and constructing a light curve, and found a bright flare, as well as quiescent emission (see Fig. 1). In subsequent observations we detected two additional flares, and quiescent emission at 4.9 and 8.5 GHz. A log of all observations, the details of the data reduction, and results can be found in Table 1 and Fig. 1. The quiescent emission at the location of the source was detected at the 2–3 σ level in individual measurements, but with a level of 2.5 and 5.5 σ for the average of all measurements at 4.9 and

8.5 GHz, respectively. Because we know the precise location of the source, these levels represent a significant detection of quiescent emission.

This set of observations comprises the first detection of persistent quiescent and flaring emission from a bona-fide brown dwarf^{1,12}. Moreover, the levels of emission—approximately 80 μ Jy in quiescence and 2 mJy at peak-flare—are highly unusual in the context of stellar coronal emission. We now seek to interpret the quiescent and flaring emission in this framework.

Coronal activity in a broad range of stars has yielded⁸ a simple empirical relation between the quiescent X-ray and radio luminosities, $L_R \approx L_X/10^{15.5} \text{ Hz}^{-1}$. Using this relation with the Chandra-determined³ upper limit on the quiescent X-ray luminosity of LP944–20, $L_X < 10^{24} \text{ erg s}^{-1}$ (0.1–10 keV band), we find that the expected quiescent radio luminosity is $L_R < 3.5 \times 10^8 \text{ erg s}^{-1} \text{ Hz}^{-1}$, which translates to a received flux density of less than 0.01 μ Jy on Earth. Therefore, the flux density we measured in quiescence is at least four orders of magnitude larger than predicted.

A similar, but slightly non-linear, relation was also found⁹ between X-ray and radio flares from a broad range of stars. The non-linearity means that for flaring emission the ratio L_X/L_R varies with the luminosity of the flare. For the X-ray luminosity detected³, $L_X \approx 1.2 \times 10^{26} \text{ erg s}^{-1}$, the relation is $L_X/L_R \approx 10^{16.5} \text{ Hz}$, and we therefore predict an observed radio flux density, $F_{\nu R} \approx 0.1 \mu\text{Jy}$. The actual observed value is approximately 2×10^4 times higher.

Even though the X-ray and radio observations were not carried

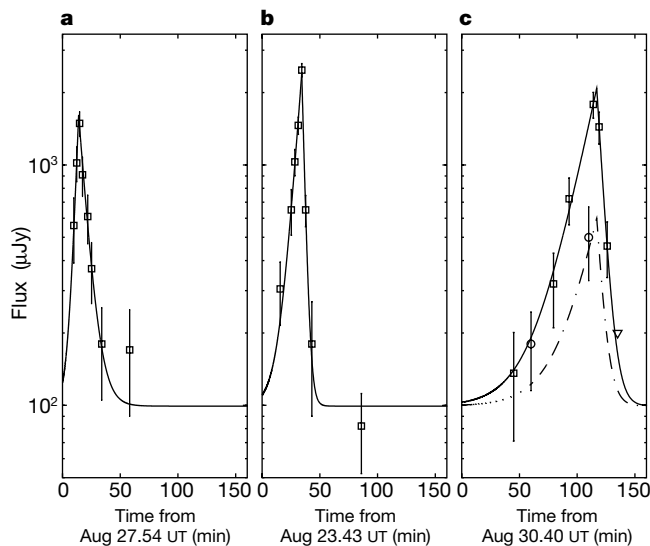


Figure 1 Light curves for the flaring radio emission from LP944–20. Data obtained at 8.5 GHz (squares) show strong flares and quiescent emission; the figure also shows self-absorbed emission at 4.9 GHz (circles; upper limit is given by an inverted triangle). The three panels are the light curves on **a**, 2000 July 27.54; **b**, 2000 August 23.43; and **c**, 2000 August 30.40. Time in minutes is relative to the starting time of each observation, and the solid line is the model fit based on the fitting equation given in Table 2. **c**, Also shown is the flux density at 4.9 GHz, which was measured simultaneously with the 8.5-GHz observation. Based on the ratio of the flux densities at these two frequencies we find that the spectral index is 2.1 ± 0.3 , which indicates that the emission at 4.9 GHz is self-absorbed. The dashed line is a model fit based on the simple model (Table 2) and the spectral shape of optically thick synchrotron emission, $F_{\nu} \propto \nu^{5/2}$. From these light curves we can calculate the magnetic field strength and the number density of trapped electrons using the equations for synchrotron emission: $(eB/m_e c) \gamma_e^2 = \nu \approx 8.5 \text{ GHz}$ and $\frac{4}{3} \sigma_T c \gamma_e^2 (B^2/8\pi) N = 4\pi d^2 \nu F_{\nu \text{ peak}} \approx 5 \times 10^{23} \text{ erg s}^{-1}$, where e and m_e are the electron charge and mass, respectively, σ_T is the Thomson cross-section, N is the total number of trapped electrons, d is the distance to the source, and $F_{\nu \text{ peak}}$ is the peak-flare flux density.

out simultaneously, X-ray flares bright enough to resolve the discrepancy are not likely. In order to conform to the Benz–Guedel relation⁹, the luminosity of the X-ray flares would have to be approximately equal to the bolometric luminosity of LP944–20. Such superflaring activity^{13,14} has been observed in nine solar-type stars, but with an average recurrence time of years to centuries. In contrast, the average recurrence time for the radio flares is of the order of several hours (see Table 2). Moreover, with such a short recurrence time, Chandra would have been expected to detect at least one of these superflares during the 13-hour observation of LP944–20. We can therefore conclude that the Benz–Guedel relation for flaring emission is severely violated, and the radio emission from LP944–20 is highly abnormal when compared to stellar emission.

The brightness temperature of the radio emission, approximately 2×10^8 K in quiescence and 4×10^9 K during flares, indicates that the emission mechanism is non-thermal. Most probably the emission is via synchrotron, which comes from relativistic electrons gyrating in a magnetic field and emitting radio waves at high harmonics of the electron gyrofrequency. Another possible mechanism is coherent emission, but the observed properties of the quiescent and flaring emission (flare timescale of the order of minutes, polarization of about 30%, broad-band emission and $T_b < 10^{10}$ K) all indicate synchrotron emission^{11,15}; for coherent emission we would expect a much shorter timescale, a much higher polarization and brightness temperature, and narrow-band emission. For a comprehensive treatment of synchrotron theory see ref. 16.

To find the spectral properties of the flaring emission, and thus the physical properties of the emission region around LP944–20, we use the observed fraction of circular polarization in the flaring emission at 8.5 GHz, $f_{\text{circ}} = 0.3 \pm 0.1$, which matches theoretical predictions^{17–19} for optically thin synchrotron emission at moderate values of the gyrating electrons' minimum Lorentz factor (γ_{min}). In addition, we find from the simultaneous observation on 2000 August 30.40 UT that the spectral index between 4.9 and 8.5 GHz was 2.1 ± 0.3 , indicating that the emission at 4.9 GHz was optically thick (see Fig. 1c). We therefore conclude that the radio spectrum of the flaring emission peaks at about 8.5 GHz, so that most of the energy in the flares at radio wavelengths is released at this frequency. Thus, we can use $\nu_{\text{peak}} = 8.5$ GHz and $F_{\nu}(8.5 \text{ GHz}) \equiv F_{\nu, \text{peak}}$ as an excellent approximation for the flare energetics (see Table 2).

Table 1 A summary of VLA observations of LP944–20, exhibiting both quiescent and flaring emission

Date (UT)	Frequency (GHz)	On-source observation time (ks)	Synthesized beam size (arcsec)	$F \pm \sigma$ (μJy)
2000 July 27.54	8.46	3.6	26.6×6.4	300 ± 45
2000 Aug 11.47	8.46	4.2	27.5×7.0	70 ± 27
2000 Aug 15.45	4.86	1.7	52.6×12.6	127 ± 53
2000 Aug 15.45	8.46	1.6	29.8×7.1	130 ± 44
2000 Aug 23.43	8.46	5.6	27.2×7.3	250 ± 26
2000 Aug 30.40	4.86	7.2	41.4×12.6	106 ± 50
2000 Aug 30.40	8.46	7.2	26.5×7.6	323 ± 49
2000 Sep 6.39	8.46	8.8	22.9×6.2	70 ± 39
2000 Sep 6.39	22.5	8.8	10.4×4.4	190 ± 170
2000 Sep 20.34	4.86	10.5	33.9×13.7	50 ± 36
2000 Sep 20.34	8.46	10.5	45.5×6.6	140 ± 45
2000 Sep 27.35	8.46	10.2	22.9×7.4	53 ± 20

The columns are (left to right): UT date of the start of each observation; observing frequency; on-source observing time; synthesized beam size; and peak flux density at the best fit position of LP944–20, with the error given as the root mean square noise on the image. Single-frequency observations consisted of several on-source scans using the VLA standard continuum mode, with the full 100-MHz bandwidth obtained in two adjacent 50-MHz bands. The flux density scale was determined using the extra-galactic source 3C48 (J0137+331), and the phase was monitored using the sources J0334–401 and J0403–360. Dual-frequency observations were done with a split array configuration in which only several antennas were used for each observing frequency. The data were reduced with the Astronomical Image Processing System (AIPS, releases 15APR1999 and 31DEC1999)²⁰. In the initial observation we detected an object with a flux $F_{8.5 \text{ GHz}} = 300 \pm 45 \mu\text{Jy}$ at $\alpha(\text{J2000}) = 03^{\text{h}} 39^{\text{m}} 35.3^{\text{s}}$, $\delta(\text{J2000}) = -35^{\circ} 25' 44.0''$, with an error of about $1''$ in both α and δ . We identify our detection with the brown dwarf LP944–20 based on the positional coincidence between this detection and the X-ray position³ provided by Chandra.

The amount of energy released in the flares is estimated by using an exponentially rising and decaying model, and solving for the peak flux density and rise and decay times (see Table 2). We find that the energy released in each of the flares at radio wavelengths is approximately the same, in the range $(2\text{--}8) \times 10^{26}$ erg. Interestingly, this energy release is several times larger than for the brightest solar flares¹⁸. From the flare model we also find that the best-fit quiescent flux density is $100 \pm 35 \mu\text{Jy}$, and that the probability of no quiescent component is 2×10^{-4} . This result serves to bolster our claim that we detected quiescent emission at a level of $100 \mu\text{Jy}$ from LP944–20.

From the fraction of circular polarization we can also estimate the value of γ_{min} using the relation for synchrotron emission $f_{\text{circ}} \approx 3/\gamma_{\text{min}}$, which gives $\gamma_{\text{min}} \approx 10$. Using the more accurate formalism of ref. 19, we get a similar result.

With the values of γ_{min} , ν_{peak} and $F_{\nu, \text{peak}}$ we now calculate the magnetic field strength and the electron density during a flare from the basic equations of synchrotron theory¹⁶, and from the minimum energy argument, which relies on the premise that the energy of a source is minimized in equipartition^{20,21}. From synchrotron theory we find that the magnetic field strength is $B \approx 5$ G, and the total number of electrons is $N_e \approx 2 \times 10^{35}$ (see Fig. 1). From equipartition we find that the “equipartition radius”²² of LP944–20 is $\theta_{\text{eq}} \approx 75 \mu\text{as}$, whereas its physical angular size is $\theta_s \approx 60 \mu\text{as}$, indicating that the energy is released from a thin shell around the source, with a volume, $V \approx 4 \times 10^{29} \text{ cm}^3$. The electrons' Lorentz factor calculated from equipartition is $\gamma_{\text{min}} \approx 10$, which is exactly the value we calculated independently from circular polarization, and hence we again find that $B \approx 5$ G and $N_e \approx 2 \times 10^{35}$. Using the volume of the emitting shell we calculate an electron number density, $n_e \approx 6 \times 10^5 \text{ cm}^{-3}$. For comparison, the surface magnetic field of Jupiter²³ is approximately 10 G, and in solar flaring emission¹⁷ the magnetic field is $B \approx 80$ G, but the electron density and energy are much lower than in the flaring emission from LP944–20.

Similarly, the magnetic fields of some active dwarf M stars^{10,11} reach strengths of a few kG, and their X-ray luminosities in quiescence^{24–26} are of order $10^{27}\text{--}10^{29} \text{ erg s}^{-1}$. These are markedly different conditions from those in and around LP944–20, as evidenced by the weak magnetic field and non-detectable quiescent X-ray emission, but they may help to clarify several lines of evidence that point to no magnetic field activity in this object. Extremely low levels of H α emission⁵, which indicate no chromospheric activity, rapid rotation⁵, which indicates no magnetic braking, and relatively old age⁴, which implies a quenched dynamo mechanism, all agree with a magnetic field strength of a few gauss rather than the much stronger fields observed in M dwarfs, the Sun and even Jupiter.

We finally note that radio synchrotron and X-ray emission are

Table 2 A summary of the model parameters characterizing each of the observed flares

Parameter	2000 July 27.5727	2000 Aug 23.4407	2000 Aug 30.4849
F_0 (μJy)	1.5 ± 0.1	2.6 ± 0.2	2.0 ± 0.2
τ_1 (min)	$3.0^{+4.2}_{-1.5}$	$6.2^{+1.5}_{-0.8}$	20 ± 5
τ_2 (min)	$5.7^{+2.4}_{-1.6}$	$1.7^{+0.4}_{-0.3}$	$5.1^{+2.2}_{-1.5}$
δt_0 (min)	± 1	± 1	± 2
E_R (10^{26} erg)	$2.0^{+1.5}_{-0.7}$	$3.1^{+0.8}_{-0.5}$	$7.7^{+2.3}_{-2.1}$

The flares were modelled using the equation: $F(t) = F_0 e^{-(t-t_0)/\tau_1} + F_q$ for $t < t_0$ and $F(t) = F_0 e^{-(t-t_0)/\tau_2} + F_q$ for $t \geq t_0$, where F_0 is the peak flux density, F_q is the quiescent flux of the source, and τ_1 and τ_2 are the rise and decay times, respectively. With this approach F_0 , t_0 , τ_1 , τ_2 and F_q are free parameters, and the total value of χ^2 for all three flares is 3.7 for nine degrees of freedom. Forcing F_q to have the same value at all times, we find that the level of quiescent emission derived from the model fit is consistent with the quiescent flux densities measured during observations of non-flaring emission; from modelling we find an average quiescent flux of $100 \pm 35 \mu\text{Jy}$, while direct observations of the quiescent emission give a value of $75 \pm 23 \mu\text{Jy}$. The energy released in the flares at radio wavelengths is calculated using the equation $E_R = 4\pi d^2 \nu_{\text{peak}} F_{\nu, \text{peak}} \tau$, where $d = 5$ pc is the distance to the source, $\nu_{\text{peak}} \approx 8.5 \times 10^9$ Hz is the peak frequency, $F_{\nu, \text{peak}}$ is the peak flux density in units of $\text{erg s}^{-1} \text{ cm}^{-2} \text{ Hz}^{-1}$, and $\tau = \tau_1 + \tau_2$ is the temporal extent of the flare. We find that all three flares released approximately the same amount of energy at 8.5 GHz.

expected in the context of a basic model of coronal emission—the magnetic reconnection process^{18,27}—and that the previously noted constancy of radio-wavelength energy release during flares appears to indicate that magnetic reconnection in LP944–20 takes place once the magnetic fields reach a critical strength, and therefore release approximately the same amount of energy every time.

It is still not clear why the emission from LP944–20 violates the Guedel–Benz relations so severely, but it is possible that this violation is tied to the difference in physical conditions in LP944–20 relative to M dwarfs, namely a very weak magnetic field. It is therefore crucial that radio and X-ray observations of this and other brown dwarfs be undertaken in order to establish whether similar conditions and unusually strong radio emission are common to some of these objects. In particular, it is clear that the radio regime is highly effective for detailed studies of both the flaring and quiescent emission because of the ease and availability of radio observations and the excellent sensitivity of radio telescopes. □

Received 26 October 2000; accepted 18 January 2001.

- Neuhäuser, R. *et al.* Search for X-ray emission from bona-fide and candidate brown dwarfs. *Astron. Astrophys.* **343**, 883–893 (1999).
- Durney, B. R., De Young, D. S. & Roxburgh, I. W. On the generation of the large-scale and turbulent magnetic fields in solar-type stars. *Sol. Phys.* **145**, 207–225 (1993).
- Rutledge, R. E., Basri, G., Martin, E. L. & Bildsten, L. Chandra detection of an X-ray flare from the brown dwarf LP944–20. *Astrophys. J.* **538**, L141–L144 (2000).
- Tinney, C. G. The intermediate-age brown dwarf LP944–20. *Mon. Not. R. Astron. Soc.* **296**, L42–L44 (1998).
- Tinney, C. G. & Reid, I. N. High-resolution spectra of very low-mass stars. *Mon. Not. R. Astron. Soc.* **301**, 1031–1048 (1998).
- Tinney, C. G. & Tolley, A. J. Searching for weather in brown dwarfs. *Mon. Not. R. Astron. Soc.* **304**, 119–126 (1999).
- Kirkpatrick, J. D., Henry, T. J. & Irwin, M. J. Ultra-cool M dwarfs discovered by QSO surveys. I: the APM objects. *Astron. J.* **113**, 1421–1428 (1997).
- Guedel, M. & Benz, A. O. X-ray/microwave relation of different types of active stars. *Astrophys. J.* **405**, L63–L66 (1993).
- Benz, A. O. & Guedel, M. X-ray/microwave ratio of flares and coronae. *Astron. Astrophys.* **285**, 621–630 (1994).
- Saar, S. H. & Linsky, J. L. The photospheric magnetic field of the dM3.5e flare star AD Leonis. *Astrophys. J.* **299**, L47–L50 (1985).
- Haisch, B., Strong, K. T. & Rodono, M. Flares on the sun and other stars. *Annu. Rev. Astron. Astrophys.* **29**, 275–324 (1991).
- Krishnamurthi, A., Leto, G. & Linsky, J. L. A search for radio emission at the bottom of the main sequence and beyond. *Astron. J.* **118**, 1369–1372 (1999).
- Schaefer, B. E., King, J. R. & Deliyannis, C. P. Superflares on ordinary solar-type stars. *Astrophys. J.* **529**, 1026–1030 (2000).
- Rubenstein, E. P. & Schaefer, B. E. Are superflares on solar analogues caused by extrasolar planets? *Astrophys. J.* **529**, 1031–1033 (2000).
- Dulk, G. A. in *LNP Vol. 291: Cool Stars, Stellar Systems and the Sun* (eds Linsky, J. L. & Stencel, R. E.) 72–82 (Berlin, Springer, 1987).
- Rybicki, G. B. & Lightman, A. P. *Radiative Processes in Astrophysics* (Wiley-Interscience, New York, 1979).
- Bruggmann, G. & Magun, A. Temporal and spectral characteristics of the circular polarization of solar microwave bursts. *Astron. Astrophys.* **239**, 347–355 (1990).
- Bastian, T. S., Benz, A. O. & Gary, D. E. Radio emission from solar flares. *Annu. Rev. Astron. Astrophys.* **36**, 131–188 (1998).
- Dulk, G. A. & Marsh, K. A. Simplified expressions for the gyrosynchrotron radiation from mildly relativistic, nonthermal and thermal electrons. *Astrophys. J.* **259**, 350–358 (1982).
- Burbidge, G. R. & Burbidge, E. M. The sources of radio emission in NGC 5128 and NGC 1316. *Astrophys. J.* **125**, 1–8 (1957).
- Readhead, A. C. S. Equipartition brightness temperature and the inverse Compton catastrophe. *Astrophys. J.* **426**, 51–59 (1994).
- Scott, M. A. & Readhead, A. C. S. The low-frequency structure of powerful radio sources and limits to departures from equipartition. *Mon. Not. R. Astron. Soc.* **180**, 539–550 (1977).
- Smith, E. J. *et al.* Jupiter's magnetic field, magnetosphere, and interaction with the solar wind—Pioneer 11. *Science* **188**, 451–455 (1975).
- Fleming, T. A., Liebert, J., Gioia, I. M. & Maccacaro, T. M dwarfs from the Einstein extended medium sensitivity survey. *Astrophys. J.* **331**, 958–973 (1988).
- Giampapa, M. S. *et al.* The coronae of low-mass dwarf stars. *Astrophys. J.* **463**, 707–725 (1996).
- Fleming, T. A., Giampapa, M. S., Schmitt, J. H. M. M. & Bookbinder, J. A. Stellar coronae at the end of the main sequence—A ROSAT survey of the late M dwarfs. *Astrophys. J.* **410**, 387–392 (1993).
- Sturrock, P. A. Chromospheric magnetic reconnection and its possible relationship to coronal heating. *Astrophys. J.* **521**, 451–459 (1999).
- Fomalont, E. Astronomical Image Processing System/AIPS. *NRAO Newsl.* **3**, 3 (1981).

Acknowledgements

We thank B. Clark for the allocation of *ad hoc* VLA time. We also thank S. R. Kulkarni, D. E. Gary and R. Sari for helpful discussions. The initial observation of LP944–20 was

undertaken as part of the National Radio Astronomy Observatory (NRAO) VLA Summer Program funded by the National Science Foundation (NSF). The National Radio Astronomy Observatory is a facility of the NSF operated under cooperative agreement by Associated Universities, Inc.

Correspondence and requests for materials should be addressed to E.B. (e-mail: ejb@astro.caltech.edu).

Strong coupling between local moments and superconducting 'heavy' electrons in UPd₂Al₃

N. K. Sato^{*†}, N. Aso[‡], K. Miyake[§], R. Shiina^{†||}, P. Thalmeier[†], G. Varelogiannis[¶], C. Geibel[†], F. Steglich[†], P. Fulde[¶] & T. Komatsubara[☆]

^{*} Department of Physics, Graduate School of Science, Nagoya University, Nagoya 464-8602, Japan

[†] Max Planck Institute for Chemical Physics of Solids, D-01187 Dresden, Germany

[‡] Neutron Scattering Laboratory, Institute for Solid State Physics, University of Tokyo, Ibaraki 319-1106, Japan

[§] Department of Physical Science, Graduate School of Engineering Science, Osaka University, Toyonaka 560-8531, Japan

[¶] Max Planck Institute for the Physics of Complex Systems, D-01187 Dresden, Germany

^{||} Institute of Electronic Structure and Laser, Foundation for Research and Technology—Hellas, 71110 Heraklion, Greece

[☆] Physics Department, Graduate School of Science, Tohoku University, Sendai 980-8578, Japan

The electronic structure of heavy-fermion compounds arises from the interaction of nearly localized 4f- or 5f-shell electrons (with atomic magnetic moments) with the free-electron-like itinerant conduction-band electrons. In actinide or rare-earth heavy-fermion materials, this interaction yields itinerant electrons having an effective mass about 100 times (or more) the bare electron mass. Moreover, the itinerant electrons in UPd₂Al₃ are found to be superconducting well below the magnetic ordering temperature^{1,2} of this compound, whereas magnetism generally suppresses superconductivity in conventional metals. Here we report the detection of a dispersive excitation of the ordered f-electron moments, which shows a strong interaction with the heavy superconducting electrons. This 'magnetic exciton' is a localized excitation which moves through the lattice as a result of exchange forces between the magnetic moments. By combining this observation with previous tunnelling measurements on this material³, we argue that these magnetic excitons may produce effective interactions between the itinerant electrons, and so be responsible for superconductivity in a manner analogous to the role played by phonons in conventional superconductors.

The hybridization or 'mixing' of atomic and conduction-electron wavefunctions in heavy-fermion compounds, together with the strong Coulomb repulsion within the atomic f shells, leads to a notable many-body effect. The atomic moments seem to progressively reduce upon lowering the temperature, and at very low temperatures, all the electrons appear to be itinerant for static physical properties—albeit with a 'heavy' effective electron (or 'quasiparticle') mass, which is much bigger than the free-electron mass. This low-temperature state is called a 'Fermi liquid', meaning

|| Present address: Department of Physics, Tokyo Metropolitan University, Tokyo 192-0397, Japan.

a degenerate Fermi gas of weakly interacting quasiparticles. However, as the temperature is reduced towards $T = 0$, an ideal Fermi liquid is almost never achieved in heavy-fermion metals. Interactions between incompletely compensated atomic moments on different sites are still present at low temperature, and may lead to a real magnetic phase transition before the 'true' Fermi-liquid behaviour is reached.

UPd_2Al_3 has a two-component electronic character that is unique among all U-based heavy-fermion superconductors, because it displays both pronounced local-moment and heavy-mass itinerant behaviour^{2,4–8}. This is due to the peculiar character of the $5f$ shell⁹ in UPd_2Al_3 , which has an average occupation of slightly less than three. Two electrons are 'localized' $5f$ electrons in the U^{4+} state, while the remaining $5f$ electron is 'itinerant' due to a larger hybridization with conduction electrons. In UPd_2Al_3 the two-component nature of the $5f$ electrons is evident from the fact that the saturation moment far below the antiferromagnetic ordering temperature T_N (≈ 14.3 K) is almost atomic-like ($0.85 \mu_B$), whereas its itinerant heavy-fermion character is clear from the Sommerfeld coefficient in the electronic specific heat, $\gamma = 140 \text{ mJ mol}^{-1} \text{ K}^{-2}$, which is much larger than for conventional metals, although nearly an order of magnitude smaller than for the most extreme heavy-fermion compounds. A large specific-heat jump, $\Delta C = 1.2\gamma T_c$, at the superconducting transition temperature ($T_c \approx 1.8$ K) signals that heavy quasiparticles condense into an anisotropic superconducting state².

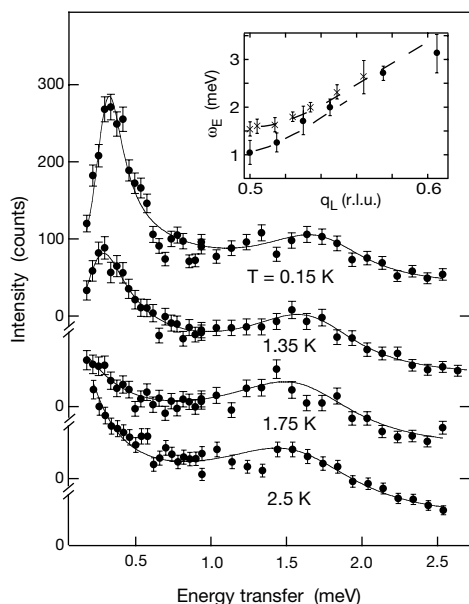


Figure 1 Temperature evolution of the inelastic neutron scattering spectrum. The experiments were performed at the IN14 spectrometer at the ILL Grenoble, and the experimental details are described elsewhere^{10–13}. The region of ω smaller than ~ 0.15 meV was inaccessible because it was masked by the strong intensity of incoherent elastic scatterings and by a tail of the antiferromagnetic (AF) Bragg peak at $\omega = 0$. The momentum transfer $\mathbf{q}$ varied along the c axis around the AF zone centre $\mathbf{Q}_0 = (0, 0, q_L = 0.5)$. Solid lines represent fits using the expression for the spectral function given in Fig. 2 legend with the assumption that α , Δ_{CEF} and J_{cf} are independent of temperature. As inferred from susceptibility²³, microscopic properties^{8,24} and our own calculation of the AF order parameter, the local component is described by a $\text{U}^{4+}(5f^2)$ CEF singlet ground state $|g\rangle$ and an excited state $|e\rangle$, probably a doublet, at an energy Δ_{CEF} (≈ 7 meV). The formation of an induced local moment in the AF regime is caused by a magnetic matrix element $\alpha = \langle e | J_{\text{cf}} | g \rangle$ (≈ 2.8) between them and $J_{\text{cf}}(\mathbf{q})$ (≈ 0.7 meV for $\mathbf{q} = \mathbf{Q}_0$). Inset, the dispersion relation of magnetic excitons at $T = 2.5$ K (crosses) and 0.15 K (circles). Dashed lines are guides to the eyes. The dispersion corresponds to an acoustic mode in the AF structure, and its finite gap at $\mathbf{q} = \mathbf{Q}_0$ is caused by uniaxial exchange anisotropies.

In Fig. 1 we show the scattering intensity as a function of the energy transfer ω ($\hbar = 1$) of neutrons for various temperatures^{10–16}. Far below T_c ($T = 0.15$ K) a double-peak structure is seen, with peak positions at $\omega \approx 0.35$ meV and 1.7 meV. We observe that the lower peak at $\omega \approx 0.35$ meV for $T = 0.15$ K is progressively reduced on increasing temperature, and turns into a quasielastic line (that is, with spectral line-shape centred around $\omega \approx 0$) above $T_c = 1.8$ K. This change through T_c suggests that the lower peak is a 'resonance' associated with the itinerant heavy quasiparticles. As the neutron-scattering cross-section is dominated by the localized $5f$ states due to their large form factor, the observation that the lower peak has high intensity suggests a sizeable interaction between localized and itinerant $5f$ electrons. We also observed that the upper peak shows little intensity variation but considerable dispersion, whereas the intensity of the lower peak is strongly reduced when the momentum transfer $\mathbf{q}$ is moving away from the antiferromagnetic zone centre. (These $\mathbf{q}$ -dependences of the spectra are not shown here, but the dispersive nature of the upper peak may be seen in the Fig. 1 inset.) Because of its large dispersion, the upper peak can be associated with a collective mode of the localized-moment system due to the intersite interaction. We now use a microscopic model to analyse this behaviour of the UPd_2Al_3 inelastic magnetic neutron scattering.

The two-component model hamiltonian is given by

$$H = \sum_{k\sigma} \epsilon_{k\sigma} c_{k\sigma}^\dagger c_{k\sigma} + \Delta_{\text{CEF}} \sum_i |e\rangle \langle e|_i - \sum_{\langle i,j \rangle} J_{\text{ff}}(ij) \mathbf{J}_i \cdot \mathbf{J}_j - J_{\text{cf}} \sum_i \mathbf{s}_i \cdot \mathbf{J}_i \quad (1)$$

The terms above describe itinerant electrons (quasiparticles) ($c_{k\sigma}$, $\mathbf{s}_i$)

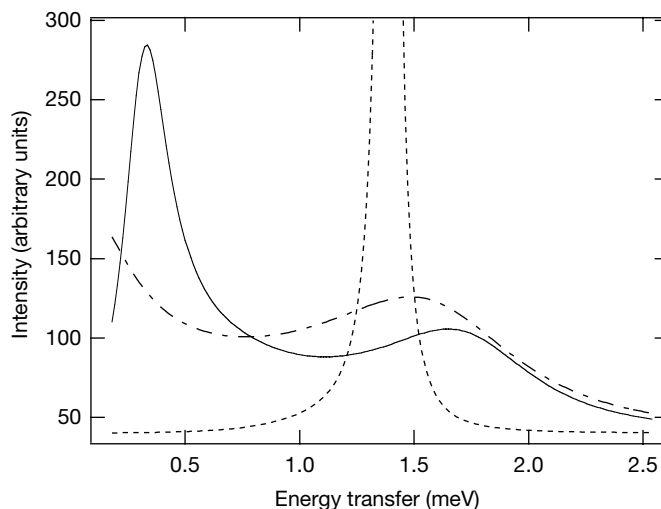


Figure 2 Qualitative evolution of the scattering intensity as a function of the coupling strength. It is proportional to $(1 + n(\omega)) \text{Im} G_{\text{ff}}$ where the $5f$ spectral function $\text{Im} G_{\text{ff}}$ is obtained from the equations of motion governed by equation (1) which lead to $G_{\text{ff}} = g_f + J_{\text{cf}}^2 g_f g_c G_{\text{ff}}$. Here g_f and g_c are the decoupled dynamical susceptibilities of the localized and itinerant $5f$ electrons, respectively; $g_f = \Delta_{\text{CEF}} \alpha^2 / (\omega^2 - \Delta_{\text{CEF}}^2)$ was approximated by $\alpha^2 / (\omega - \Delta_{\text{CEF}})$ for $\omega \geq 0$ to obtain better convergence in the fitting procedure, and g_c is the damped oscillator function¹³. Dotted line, weak-coupling case ($J_{\text{cf}}^2 = 0.005/\chi_{\text{st}}$, where χ_{st} is the static susceptibility of the quasiparticles); only a slightly broadened magnetic-exciton peak at 1.4 meV appears. Dashed-dotted line, strong-coupling case ($J_{\text{cf}}^2 = 0.12/\chi_{\text{st}}$) for $T \approx T_c$; both a strongly broadened inelastic peak around 1.5 meV and a quasielastic contribution are visible. Owing to this broadening effect the CEF splitting is difficult to observe experimentally²⁴. Solid line, strong-coupling case ($J_{\text{cf}}^2 = 0.12/\chi_{\text{st}}$) for $T = 0$; the superconducting-gap-opening effect leads to the double-peak spectrum.

with energies $\epsilon_{k\sigma}$, crystalline-electric-field (CEF) splitting (Δ_{CEF}) of localized $5f$ electrons (J_l), an effective inter-site exchange interaction between the latter (J_{ff}) and an on-site coupling of itinerant and localized electrons (J_{cf}), respectively. (The CEF means potentials arising from neighbouring ions.) If the CEF states of U^{4+} ions were isolated from each other (that is, $J_{ff} = 0$), we would observe a sharp peak in the spectrum of scattered neutrons at an energy transfer $\omega = \Delta_{\text{CEF}}$ caused by the local CEF excitation to an excited state $|e\rangle$. However, the effective exchange between the local moments will lead to a band of magnetic excitons whose dispersion is essentially given by:

$$\omega_E(\mathbf{q}) = \Delta_{\text{CEF}} - \alpha^2 J_{ff}(\mathbf{q}) \quad (2)$$

Here α is a matrix element defined in the Fig. 1 legend. These collective modes can be interpreted as local CEF excitations propagating from site to site. In the regime of interest, $T \ll \Delta_{\text{CEF}}$ holds and the magnetic excitons obey Bose statistics, similar to lattice vibrations (phonons).

The two-component $5f$ -electron character implies that there is a strong exchange coupling J_{cf} (last term in equation (1)) between the quasiparticles and localized electrons. This couples magnetic

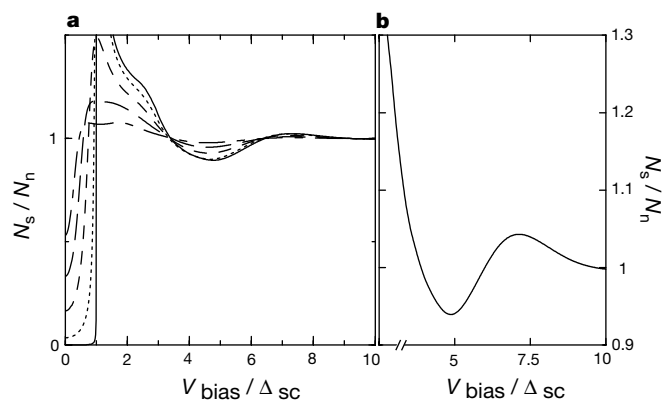


Figure 3 Tunnelling density of states calculated as function of the applied voltage V_{bias} for various temperatures. N_s/N_n is obtained from the analytic continuation of the gap function $\Phi_0(i\omega_n)$ which is the solution of the first Eliashberg equation:

$$\Phi_0(i\omega_n) = -T \sum_{\mathbf{k}'} \sum_m V_0(i\omega_n - i\omega_m) \frac{d_0^2(\mathbf{k}') \Phi_0(i\omega_m)}{\xi_{\mathbf{k}'}^2 + \omega^2(\mathbf{k}', i\omega_m) + \Phi^2(\mathbf{k}', i\omega_m)}$$

Here $d_0(\mathbf{k})$ is the $\mathbf{k}$ -dependent form factor of the anisotropic gap function in equation (3), V_0 is the amplitude of the corresponding pair potential, $\Phi(\mathbf{k}, i\omega_m) = d_0(\mathbf{k}) \Phi_0(i\omega_m)$ and $\omega(\mathbf{k}, i\omega_n)$ is determined by a similar second Eliashberg equation. We solve the equation for $\Phi_0(i\omega_n)$ by treating the momentum and frequency dependences separately. In this approximation the form factor does not have a major influence on the frequency dependence of $\Phi_0(i\omega_n)$, and we can take $\xi_{\mathbf{k}} \approx \xi$ and $d_0^2(\mathbf{k}) \approx \langle d_0^2(\mathbf{k}) \rangle_{\text{FS}}$ (FS, Fermi surface) which is a constant smaller than one. The constant is absorbed into V_0 , which is adjusted to reproduce the observed critical temperature T_c . Within this approximation we can only obtain a momentum-averaged gap function. Because only the square $d_0^2(\mathbf{k})$ enters above, our analysis is equivalent to conventional Eliashberg theory. This is justifiable, because we compare our results with tunnelling data along the c axis³ that have no contribution from the node regions, which lie at the intersection of $\mathbf{k} = \pm 0.5 \mathbf{Q}_0$ (planes $\perp c$ -axis) with the Fermi surface. **a**, The various lines correspond to different temperatures (full line, $T = 0.25 T_c$; dotted, $0.6 T_c$; dashed, $0.8 T_c$; long-dashed, $0.9 T_c$; dot-dashed, $0.95 T_c$). The dip structure and the relatively weak sensitivity to temperature of the peak position at the superconducting-gap edge agrees well with the differential tunnel-conductance results of ref. 3. The shoulder in the $T = 0.25 T_c$ density of states at $2.5 \Delta_{\text{SC}}$ was not observed³, presumably due to experimental resolution. **b**, Results for $T = 0.25 T_c$, normalized to the experimental low- T energy-dependent density of states in the normal state, N_n , which shows a shallow minimum at the Fermi energy.

excitons to particle-hole excitations and will lead to a shifting and a broadening of their spectrum. This coupling is manifest in the equation of motion for the f electrons, from which the observable $5f$ -excitation spectrum may be obtained as briefly described in the Fig. 2 legend¹⁷. Before applying the model analysis to our data, we show its principal results schematically in Fig. 2. For small J_{cf} we predict only one peak centred at $\omega_E(\mathbf{q})$, which has acquired a finite width (dotted line). An increase of J_{cf} by a factor of five leads to a strong damping of this upper peak, and there appears an additional low-energy quasielastic peak (dash-dotted line). Below T_c , this quasielastic peak has been shifted to a finite energy and appears as a 'resonance' peak, caused by the opening of the superconducting gap (full line). Marked effects of superconductivity on the spin response are also known from the high- T_c compounds, where a pronounced resonance peak at 41 meV appears below T_c (ref. 18). Its detailed physical origin is, however, different: the position of the peak is independent of temperature and its intensity vanished when approaching T_c , contrary to the present case where the resonance mode softens on $T \rightarrow T_c$ and is still present above T_c as a quasielastic line. This is possibly connected with the fact that in UPd_2Al_3 the resonance appears directly in the particle-hole channel, whereas in the high- T_c case it is in the two-particle channel and can only be seen as long as it can couple to the condensate (see ref. 19 and references therein).

A quantitative analysis of our experimental data using our model leads to the calculated curves for the spectra in Fig. 1. The neutrons couple dominantly to the local moments; therefore in the case $J_{cf} = 0$ only the magnetic excitons at $\omega_E(\mathbf{q})$ would lead to a scattering, but not the itinerant excitations centred at the superconducting-gap energy denoted by $2\Delta_{\text{SC}}$. However, for non-zero J_{cf} and under the condition that the magnetic-exciton and the superconducting-gap energies are nearly degenerate, both localized and itinerant excitations appear in a double-peak spectrum because of a weight transfer from the former to the latter. The observed broadened peaks are strongly shifted from the original nearly equal positions $\omega_E(\mathbf{q})$ and $2\Delta_{\text{SC}}$ due to their strong mutual repulsion caused by J_{cf} . We stress that, therefore, the lower-energy peak is related to the superconducting-gap opening, but is not a direct measure of $2\Delta_{\text{SC}}$.

In Fig. 1 inset we show the obtained magnetic-exciton dispersion $\omega_E(\mathbf{q}, T)$, and conclude that at $T = 0.15 \text{ K}$, $\omega_E(\mathbf{Q}_0) \approx 1.05 \pm 0.25 \text{ meV}$ is only slightly larger than $2\Delta_{\text{SC}} \approx 6 k_B T_c$, which is also obtained from the fit and agrees well with NMR results²⁰. This confirms the near degeneracy of the magnetic-exciton and superconducting-gap energies, which is important not only for the marked effect of J_{cf} on the spectrum of interacting localized and itinerant modes, but also for the unusual behaviour in the tunnelling spectra of ref. 3, as will be shown later.

Our model implies that the pairing potential $V(\mathbf{k}, \mathbf{k}')$ between the quasiparticles should be proportional to $\omega_E(\mathbf{k} - \mathbf{k}')^{-1}$. We point out that the superconducting-gap energy as obtained above is an averaged value of $\Delta_{\text{SC}}(\mathbf{k})$ over all directions in $\mathbf{k}$ -space, while it differs from the value deduced from the tunnelling experiments, measured along the crystallographic c axis. This indicates a substantial anisotropy of the gap function. On the other hand, the existence of the strong low-energy magnetic response at the antiferromagnetic zone centre $\mathbf{k} = \mathbf{Q}_0$ for very low temperatures in the superconducting state (see Fig. 1) implies the symmetry relation $\Delta_{\text{SC}}(\mathbf{k} + \mathbf{Q}_0) = -\Delta_{\text{SC}}(\mathbf{k})$: only then a non-vanishing coherence factor in the expression for the dynamical magnetic susceptibility is assured^{13,21}. This argument is similar to that in the high- T_c case mentioned above¹⁸. The most obvious even-parity unconventional gap function which exhibits this symmetry would be:

$$\Delta_{\text{SC}}(\mathbf{k}) = \Delta_0 f_{\Gamma_1}(\mathbf{k}) \cos k_z c \quad (3)$$

Here Δ_0 is the amplitude of the gap function, c is the lattice constant

along the c axis, and $f_{\Gamma}(\mathbf{k})$ is a fully symmetric representation (or linear combination of such representations) of the symmetry group D_{6h} . Then the line nodes of $\Delta_{SC}(\mathbf{k})$ are given as the intersections of $\mathbf{k} = \pm 0.5 \mathbf{Q}_0$ with the Fermi surface.

Finally, we compare our neutron-scattering spectra with those of the tunnelling measurements. Within the context of a separable potential approximation, the anisotropic Eliashberg equation may be approximated by the conventional one, as explained in the Fig. 3 legend, provided that we replace the square of the form factor $d_0(\mathbf{k}) = f_{\Gamma}(\mathbf{k}) \cos k_z c$ by a momentum-averaged constant and adjust the potential to obtain the correct T_c . This approximation is also justified because the tunnelling current is observed along the c axis, and in this case the node regions do not contribute. From this conventional Eliashberg calculation, using the observed spectrum, we obtain an averaged gap ratio $2\Delta_{SC}/k_B T_c = 5.6$: this is very close to the above-mentioned value $2\Delta_{SC}/k_B T_c \approx 6$, which was independently obtained from the neutron scattering results. Therefore, our boson- and gap-energy estimates are consistent with the critical temperature within the Eliashberg framework.

Our analysis is further supported by the fact that the same calculations reproduce the distinct 'dip and second-peak structure' found in the tunnelling spectrum around 1 meV; compare the calculated spectrum shown in Fig. 3a and b with the observed spectrum given in Fig. 4 of ref. 3. We stress that our calculation also reproduces the anomalous, only weakly temperature-dependent, behaviour of the peak position related to the superconducting gap, as shown in Fig. 3 of ref. 3. This type of unusual tunnelling spectrum arises within the Eliashberg theory precisely when the superconducting gap is comparable to the energy of the boson that mediates the superconducting pairing²². Therefore, our analysis strongly suggests that the bosons identified by the neutron scattering experiments are responsible for superconductivity. $\square$

Received 5 June 2000; accepted 3 January 2001.

- Geibel, C. *et al.* Heavy-fermion superconductivity at $T_c = 2$ K in the antiferromagnet UPd_2Al_3 . *Z. Phys. B* **84**, 1–2 (1991).
- Steglich, F. *et al.* in *Proc. Physical Phenomena at High Magnetic Fields-II* (eds Fisk, Z., Gor'kov, L., Meltzer, D. & Schrieffer, R.) 125–138 (World Scientific, Singapore, 1996).
- Jourdan, M. *et al.* Superconductivity mediated by spin fluctuations in the heavy-fermion compound UPd_2Al_3 . *Nature* **398**, 47–49 (1999).
- Caspary, R. *et al.* Unusual ground state properties of UPd_2Al_3 : Implications for the coexistence of heavy fermion superconductivity and local moment antiferromagnetism. *Phys. Rev. Lett.* **71**, 2146–2149 (1993).
- Feyerherm, R. *et al.* Coexistence of local moment magnetism and heavy fermion superconductivity in UPd_2Al_3 . *Phys. Rev. Lett.* **73**, 1849–1852 (1994).
- Takahashi, T. *et al.* Dual character of 5f electrons in UPd_2Al_3 observed by high-resolution photoemission spectroscopy. *J. Phys. Soc. Jpn* **65**, 156–159 (1996).
- Knöpfle, K. *et al.* The Fermi surface of UPd_2Al_3 . *J. Phys. Condens. Matter* **8**, 901–909 (1996).
- Yaouanc, A. *et al.* Study of the uranium 5f shell in UPd_2Al_3 and URu_2Si_2 by the x-ray magnetic circular dichroism technique. *Phys. Rev. B* **58**, 8793–8799 (1998).
- Yotsushashi, S., Kusunose, H. & Miyake, K. Orbital dependence in Kondo effect enlarged by Hund's-rule coupling. *J. Phys. Soc. Jpn* **70**, 186–191 (2001).
- Sato, N. *et al.* Spin fluctuations in the heavy fermion superconductor UPd_2Al_3 studied by neutron inelastic scattering. *J. Phys. Soc. Jpn* **66**, 1884–1887 (1997).
- Sato, N. *et al.* Possible spin-fluctuation mediated superconductivity in UPd_2Al_3 . *J. Phys. Soc. Jpn* **66**, 2981–2984 (1997).
- Sato, N. *et al.* Observation of two low-energy responses in UPd_2Al_3 —interaction of magnetism and superconductivity? *J. Alloys Comp.* **271–273**, 433–436 (1998).
- Bernhoeft, N. *et al.* Enhancement of magnetic fluctuations on passing below T_c in the heavy fermion superconductor UPd_2Al_3 . *Phys. Rev. Lett.* **81**, 4244–4247 (1998).
- Mason, T. E. & Aeppli, G. Neutron scattering studies of heavy fermion systems. *Mat.-fys. Medd.* **45**, 231–245 (1997).
- Metoki, N. *et al.* Superconducting energy gap observed in the magnetic excitation spectra of a heavy fermion superconductor UPd_2Al_3 . *Phys. Rev. Lett.* **80**, 5417–5420 (1998).
- Hayden, S. Excitations in exotic superconductors. *Phys. World* **12**(2), 18 (1999).
- Buyers, W. J. L. & Holden, T. M. in *Handbook on the Physics and Chemistry of the Actinides* (eds Freeman, A. J. & Lander, G. H.) 239–327 (Elsevier, Amsterdam, 1985).
- Fong, H. F. *et al.* Phonon and magnetic neutron scattering at 41 meV in $\text{YBa}_2\text{Cu}_3\text{O}_7$. *Phys. Rev. Lett.* **75**, 316–319 (1995).
- Demler, E. *et al.* π excitation of the t-J model. *Phys. Rev. B* **58**, 5719–5730 (1998).
- Kyongaku, M. *et al.* NMR and NQR studies of magnetism and superconductivity in the antiferromagnetic heavy fermion superconductors UM_2Al_3 ($M = \text{Ni}$ and Pd). *J. Phys. Soc. Jpn* **62**, 4016–4030 (1993).
- Bernhoeft, N. Superconductor order parameter symmetry in UPd_2Al_3 . *Eur. Phys. J. B* **13**, 685–694 (2000).

- Varelogiannis, G. On the limits of consistency of Eliashberg theory and the density of states of high- T_c superconductors. *Z. Phys. B* **104**, 411–422 (1997).
- Grauel, A. *et al.* Tetravalency and magnetic phase diagram in the heavy-fermion superconductor UPd_2Al_3 . *Phys. Rev. B* **46**, 5818–5821 (1992).
- Schenck, A. *et al.* Study of the positive muon Knight shift in UNi_2Al_3 : evidence for a tetravalent U^{4+} -state and crystalline electric field (CEF) splitting. *Eur. Phys. J. B* **13**, 245–256 (2000).
- Krimmel, A. *et al.* Magnetic excitations and the search for crystal-field transitions in the heavy-fermion superconductor UPd_2Al_3 . *J. Phys. Condens. Matter* **8**, 1677–1685 (1996).

Acknowledgements

Neutron scattering experiments were performed in collaboration with G. H. Lander, B. Roessli, N. Bernhoeft, A. Hiess, Y. Endoh, and with the help of the cryogenic and technical staffs of ILL, Grenoble. We thank O. Sakai, M. Lang, N. Bernhoeft, G. H. Lander and T. Dahm for discussions and comments, and D. Preston and M. Grosche for a critical reading of the manuscript. N. K. S. is supported by a Grant-in-Aid from the Ministry of Education, Science, Sports and Culture, Japan, and K. M. is supported by the COE project of Monbusho, Japan.

Correspondence and requests for materials should be addressed to N.K.S. (e-mail: kensho@edu3.phys.nagoya-u.ac.jp).

Loss of superconductivity with the addition of Al to MgB_2 and a structural transition in $\text{Mg}_{1-x}\text{Al}_x\text{B}_2$

J. S. Slusky, N. Rogado, K. A. Regan, M. A. Hayward, P. Khalifah, T. He, K. Inumaru, S. M. Loureiro, M. K. Haas, H. W. Zandbergen & R. J. Cava

Department of Chemistry and Princeton Materials Institute, Princeton University, Princeton, New Jersey 08544, USA

The basic magnetic and electronic properties of most binary compounds have been well known for decades. The recent discovery¹ of superconductivity at 39 K in the simple binary ceramic compound magnesium diboride, MgB_2 , was therefore surprising. Indeed, this material has been known and structurally characterized since the mid 1950s (ref. 2), and is readily available from chemical suppliers (it is commonly used as a starting material for chemical metathesis reactions³). Here we show that the addition of electrons to MgB_2 , through partial substitution of Al for Mg, results in the loss of superconductivity. Associated with the Al substitution is a subtle but distinct structural transition, reflected in the partial collapse of the spacing between boron layers near an Al content of 10 per cent. This indicates that superconducting MgB_2 is poised very near a structural instability at slightly higher electron concentrations.

Electronic structure calculations on MgB_2 have shown that the states at the Fermi energy (E_F) are primarily derived from boron orbitals^{4–6}. Doping on the Mg site is therefore expected to introduce electrons into the bands at E_F with relatively little disruption of the electronic network. Solid solutions of the type $\text{Mg}_{1-x}\text{Al}_x\text{B}_2$ were therefore synthesized by direct reaction of the elements to test the effect of electron concentration on the superconducting transition temperature, T_c . AlB_2 is the prototype compound for this structure type, and the $\text{Mg}_{1-x}\text{Al}_x\text{B}_2$ series has been previously reported⁷ but not physically characterized. Starting materials were bright Mg flakes (Aldrich), fine Al powder (Alfa Inorganics), and sub-micrometre amorphous B powder (Callery Chemical). Owing to the poor reactivity of crystalline boron at low temperatures, and MgO contamination of fine Mg powders, the selection of starting materials is important. Starting materials were lightly mixed in half-gram batches, and pressed into pellets. The pellets were placed on Ta foil, which was in turn placed on Al_2O_3 boats, and fired in a tube furnace

under a mixed gas of 95% Ar + 5% H₂. The samples were heated for 1 hour at 600 °C, 1 hour at 800 °C, and 1 hour at 900 °C. After cooling, they were pressed into pellets and fired for an additional 2 hours at 900 °C, then quickly cooled to room temperature. The resulting pellets have very low densities, making them unsuitable for resistivity measurements. Identical results to those reported here were obtained for materials synthesized by heating in sealed evacuated Ta tubes at 925 °C for 4 hours.

The materials prepared by the above method in the composition region of interest for study of the superconducting properties—Mg_{1-x}Al_xB₂ with 0 ≤ *x* ≤ 0.40—were all found to be solely of the AlB₂ type by powder X-ray diffraction. This crystal structure consists of honeycomb-net planes of boron, separated by triangular planes of the metals. MgB₄, the product expected for decomposition of MgB₂ due to Mg volatility at elevated temperatures⁸, was not present. The materials were single-phase AlB₂ type in the lower and the higher concentration ranges studied. In the middle region, however, two-phase mixtures of two different AlB₂-type phases were observed, distinguished by significantly different *c* lattice parameters. Partial representation of the powder diffraction patterns showing the (002) and (110) diffraction lines, characteristic of the inter-plane boron sheet separation and the in-plane boron net size, are shown in Fig. 1. The (002) reflection is sharp at both low and high Al concentrations, but becomes two broad peaks at intermediate concentrations. The ratio of those peak intensities changes with Al concentration in a manner characteristic of the mixing of two phases. Therefore the materials are single phase for *x* up to approximately 0.1, and again beyond *x* = 0.25. The incompatibility of the two phases is strong enough in character to result in a distinct two-phase region between them. Elemental analysis by energy dispersive X-ray analysis in a transmission electron microscope (Philips CM200) indicated that the particles contained a bimodal distribution of Al concentrations at intermediate Al contents, even within very small particles (0.1 μm in some cases). The presence of both Al-rich and Mg-rich phases in the composition region near *x* = 0.2 gives rise to the double-peak (002) structure and its dependence on concentration. A distribution of Al contents within those phases gives rise to the generally broadened (002) reflections observed in this region. The relatively smaller shift of the (110) reflection indicates that the in-plane size of the boron net does not change as dramatically as the inter-plane spacing at this structural instability.

The composition dependence of the crystallographic cell parameters was obtained by least-squares fitting to the positions of nine

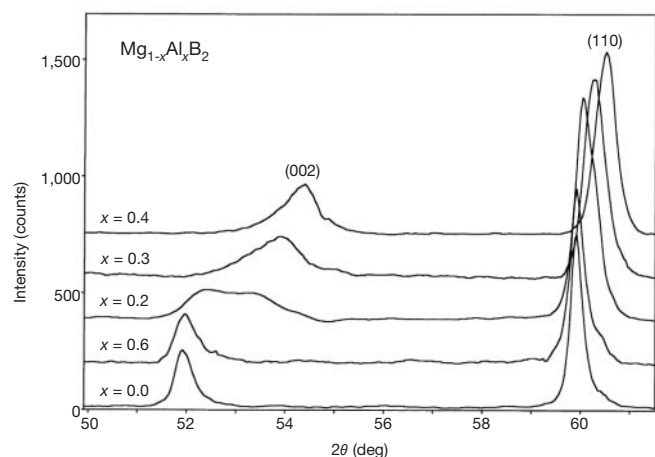


Figure 1 Changes in powder X-ray diffraction patterns with varying Al concentrations in Mg_{1-x}Al_xB₂. The figure shows diffraction patterns (Cu Kα radiation) in the region of the (002) and (110) reflections for selected compositions in the vicinity of the phase separation.

powder diffraction peaks between 2θ values of 20° and 90°. The results are presented in Fig. 2. The in-plane lattice parameter (*a*), which characterizes the dimension of the boron honeycomb net, is relatively insensitive to Al concentration. Within the large family of metal diborides⁹, the in-plane lattice parameters are relatively constant, reflecting the rigidity of the boron honeycomb layer. The separation between boron planes (the *c* axis), on the other hand, changes significantly with Al substitution. There is a discontinuity in the *c*-axis length on passing through the two-phase region centred near *x* = 0.2. In addition, *c* varies differently with aluminium concentration within the two AlB₂ type phases (that is, for 0 ≤ *x* ≤ 0.10 and *x* ≥ 0.25). The change in *c* with Al concentration is significantly greater in the phase with high Al content. This suggests the presence of a difference in electron orbital filling at low and high Al concentrations which is strongly coupled to the spacing between the boron layers.

The temperature-dependent magnetization at low temperatures was measured for these materials, in the form of loose powders, in a Quantum Design PPMS magnetometer under an applied d.c. field of 15 Oe. The data, taken on heating after cooling in the absence of a field, are presented in Fig. 3. MgB₂ synthesized by the method described here has a very sharp superconducting transition, with a *T_c* of 38 K. The transition is apparently one degree lower than that reported¹⁰ for samples synthesized with isotopically pure ¹¹B, but is similarly narrow in width, and much sharper than that found for commercial powder¹⁰. Examination by electron microscopy indicated that the particle sizes for the Mg_{1-x}Al_xB₂ powders were very small, of the order of a few tenths of a micrometre, and further that the two phases present for compositions in the two-phase region were intergrown on a nanometre length scale. Therefore the granular nature of the superconducting phase and possible proximity effects are expected to influence the shapes of the curves in Fig. 3, as is the presence of a distribution of Al contents in the powders. *T_c* drops by a few degrees with electron/Al doping for Al contents less than 0.10 (from 38 K at *x* = 0 to 36 K at *x* = 0.1), and the transitions remain sharp. The transitions broaden significantly for Al concentrations greater than *x* = 0.10. The X-ray and magnetic data suggest that *x* = 0.1 is barely within the two-phase region.

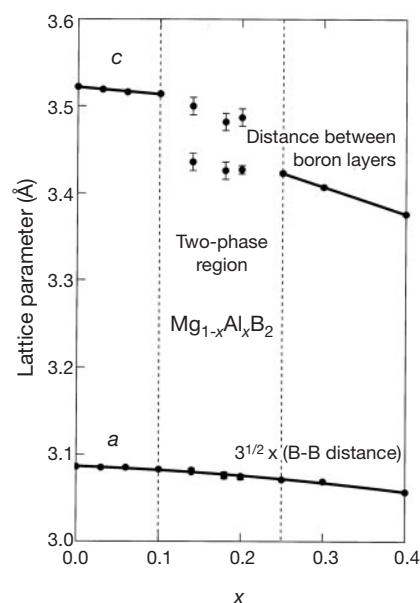


Figure 2 Partial collapse of the spacing between the boron layers in Mg_{1-x}Al_xB₂. The figure shows variation of the in-plane (*a*) and between-plane (*c*) lattice parameters as a function of aluminium concentration. In the two-phase region, *c*-axis values for both phases are shown.

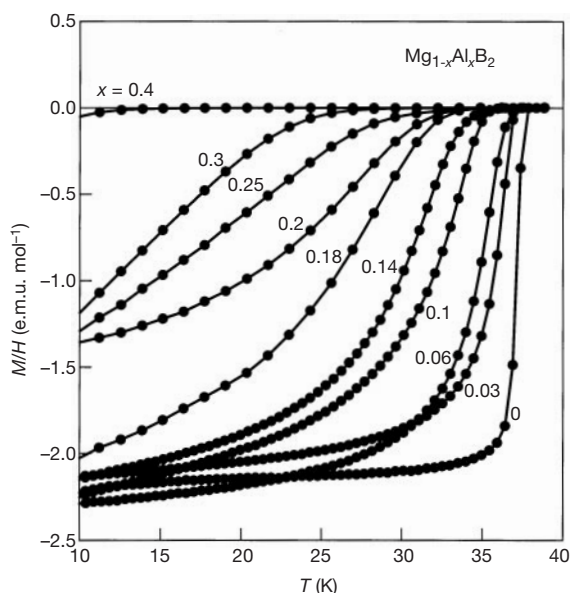


Figure 3 Magnetization of $\text{Mg}_{1-x}\text{Al}_x\text{B}_2$ as a function of temperature and Al concentration. The figure shows magnetic characterization of the superconducting transitions for polycrystalline powders of $\text{Mg}_{1-x}\text{Al}_x\text{B}_2$ for $x = 0-0.40$. A field of 15 Oe was applied after cooling the samples in zero field.

The onsets of the transitions are all in the vicinity of 36 K, the transition temperature for $x = 0.10$. There is no bulk superconducting transition for $x = 0.3$ and $x = 0.4$. The data are interpreted as indicating that bulk superconductivity in single-phase material is suppressed for x greater than 0.10. The superconducting transitions observed in the two-phase region are due to both the decreasing amount of the bulk superconducting phase with increasing Al content, and also to the distribution of Al contents in the powders. The structural and magnetic data taken together therefore indicate that the disappearance of bulk superconductivity in $\text{Mg}_{1-x}\text{Al}_x\text{B}_2$ occurs at the same Al concentrations at which a structural transition occurs—this transition results in the partial collapse of the separation between boron planes.

The similarity of the calculated electronic density of states (DOS) for MgB_2 and AlB_2 (ref. 6) indicates that the effect of substituting Al for Mg can be considered primarily as a simple filling of available electronic states, with one electron donated per Al, within a 'rigid band' picture. The calculations show⁴⁻⁶ that there is a sharp drop in the DOS of MgB_2 at only slightly higher electron concentrations. The gradual decrease of T_c from 38 to 36 K with increasing Al content in the single-phase region below $x = 0.1$ may therefore be due to the expected decrease of the DOS at E_F with increasing electron count, consistent with a conventional origin for the superconductivity¹⁰.

High-resolution structural study of uniform powders, if they prove possible to synthesize, would help to determine how the superconductivity and the structural transition are linked on an electronic level. The proximity of MgB_2 to a structural instability is consistent with a general picture for conventional high-temperature superconductors as being near chemical phase instability owing to strong electron–lattice coupling. However, our data for $\text{Mg}_{1-x}\text{Al}_x\text{B}_2$ do not suggest that the structural instability in this case is driven by competition between superconductivity and a structural distortion that would decrease the density of electronic states at the Fermi energy; this is because the Al substitution both decreases T_c and decreases the DOS at E_F before the transition occurs. It is at first sight surprising that the observed structural instability in MgB_2 involves a partial collapse of the separation of the boron layers, not a change in the boron–boron in-plane distance. It will be of interest

to determine whether this collapse is associated with special characteristics of the electronic band filling near 0.1 excess electrons per cell. □

Received 8 February; accepted 16 February 2001.

1. Nagamatsu, J., Nakagawa, N., Muranaka, T., Zenitani, Y. & Akimitsu, J. Superconductivity at 39 K in magnesium diboride. *Nature* **410**, 63–64 (2001).
2. Jones, M. E. & Marsh, R. E. The preparation and structure of magnesium boride, MgB_2 . *J. Am. Chem. Soc.* **76**, 1434–1436 (1954).
3. Killian, E. G. & Kaner, R. B. Synthesis of refractory ceramics via rapid metathesis reactions between solid-state precursors. *Chem. Mater.* **8**, 333–343 (1996).
4. Kortis, J., Mazin, I. I., Belaschenko, K. D., Antropov, V. P. & Boyer, L. L. Superconductivity of metallic boron in MgB_2 . Preprint cond-mat/0101446 at (<http://xxx.lanl.gov>) (2001).
5. Armstrong, D. R. & Perkins, P. G. Electronic band structure of magnesium diboride. *J. Chem. Soc. Faraday Trans. 2* **75**, 12–16 (1979).
6. Ivanovskii, A. L. & Medvedeva, I. Interatomic interactions and electronic structure of hexagonal magnesium, aluminum and silicon diborides: Ab initio full-potential LMTO calculations. *Russ. J. Inorg. Chem.* **45**, 1234–1240 (2000).
7. Vekshina, N. V., Markovskii, L. Ya., Kondrashev, Yu. D. & Voevodskaya, T. K. Binary borides of aluminum and magnesium. *Zh. Prikl. Khim.* **44**, 958–963 (1971).
8. Massalski, T. B. (ed.) *Binary Alloy Phase Diagrams* 498–499 (ASM International, Materials Park, Ohio, 1990).
9. Villars, P. & Calvert, L. D. *Pearson's Handbook of Crystallographic Data for Intermetallic Phases* 127–128 (ASM International, Materials Park, Ohio, 1991).
10. Bud'ko, S. L. *et al.* Boron isotope effect in superconducting MgB_2 . Preprint cond-mat/0101463 at (<http://xxx.lanl.gov>) (2001).

Acknowledgements

This work was supported by the US Department of Energy and the National Science Foundation.

Correspondence and requests for materials should be addressed to R.J.C. (e-mail: rcava@princeton.edu).

Electrical spin injection and accumulation at room temperature in an all-metal mesoscopic spin valve

F. J. Jedema, A. T. Filip & B. J. van Wees

Department of Applied Physics and Materials Science Centre, University of Groningen, Nijenborgh 4.13, 9747 AG Groningen, The Netherlands

Finding a means to generate, control and use spin-polarized currents represents an important challenge for spin-based electronics¹⁻³, or 'spintronics'. Spin currents and the associated phenomenon of spin accumulation can be realized by driving a current from a ferromagnetic electrode into a non-magnetic metal or semiconductor. This was first demonstrated over 15 years ago in a spin injection experiment⁴ on a single crystal aluminium bar at temperatures below 77 K. Recent experiments⁵⁻⁸ have demonstrated successful optical detection of spin injection in semiconductors, using either optical injection by circularly polarized light or electrical injection from a magnetic semiconductor. However, it has not been possible to achieve fully electrical spin injection and detection at room temperature. Here we report room-temperature electrical injection and detection of spin currents and observe spin accumulation in an all-metal lateral mesoscopic spin valve, where ferromagnetic electrodes are used to drive a spin-polarized current into crossed copper strips. We anticipate that larger signals should be obtainable by optimizing the choice of materials and device geometry.

Spin accumulation plays an important role in spin-polarized

transport phenomena such as giant magnetoresistance (GMR), domain-wall magnetoresistance and spin-current-induced magnetization switching experiments^{9–12}. We wanted to study the effect isolated from other spin-related phenomena, such as spin-dependent interface scattering, anisotropic magnetoresistance (AMR) and Hall effects. Large spin-accumulation effects have been claimed, using a ferromagnet–normal-metal–ferromagnet sandwich geometry^{13,14}. However, the interpretation of the data has raised problems, because the magnitude of the observed effect requires a spin polarization of the current in the normal metal of around 100% (refs 15–18). Another problem is that the magnetoresistance effects of the ferromagnetic contacts, such as AMR and Hall effects, can mask or even mimic the spin-accumulation signal.

We have fabricated mesoscopic lateral spin valves to isolate the spin accumulation signal completely, using a multi-terminal geometry. In our spin-injection experiments, we use permalloy $\text{Ni}_{80}\text{Fe}_{20}$ (Py) electrodes to drive a spin-polarized current into copper (Cu) cross strips (Fig. 1). We observe clear spin valve signals at $T = 4.2$ K, as well as at room temperature. From our analysis we deduce a spin-flip length λ_N in the Cu wire of about $1\text{ }\mu\text{m}$ at 4.2 K, which is reduced to about 350 nm at room temperature.

Two batches of samples were made in a two-step lift-off process, using electron-beam lithography for patterning. The first batch (1) had a fixed Py electrode spacing L of 250 nm , and in the second batch (2) L is varied from 250 nm to $2\text{ }\mu\text{m}$. To avoid magnetic fringe fields from the ferromagnetic electrodes, the 40-nm -thick Py electrodes (Py1, Py2) were sputter deposited first on a thermally oxidized silicon substrate. Different geometric aspect ratios of Py1 and Py2 are used to obtain different coercive fields¹⁹. We can thus control the relative magnetization configuration (parallel/anti-parallel) of Py1 and Py2, by sweeping an applied magnetic field B ,

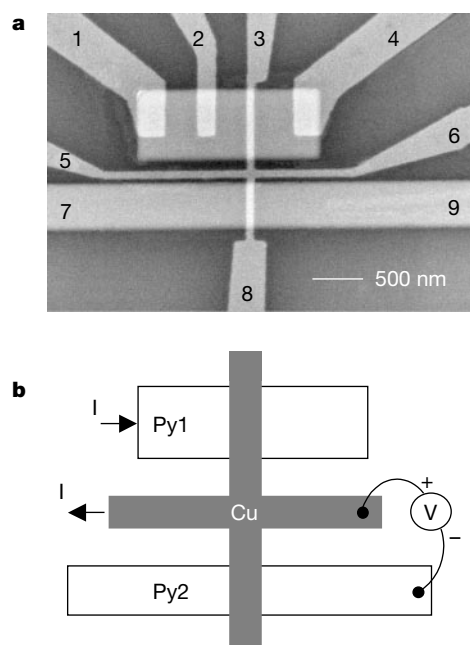


Figure 1 Sample layout. **a**, Scanning electron microscope image of the mesoscopic spin valve junction. The two wide horizontal strips are the ferromagnetic electrodes Py1 and Py2. The vertical arms of the Cu cross (contacts 3 and 8) lie on top of the Py strips; the horizontal arms of the Cu cross form contacts 5 and 6. Contacts 1, 2, 4, 7 and 9 are attached to Py1 and Py2 to allow four terminal AMR measurements of the Py electrodes. **b**, Schematic representation of the non-local measurement geometry. Current is entering from contact 1 and extracted at contact 5. The voltage is measured between contact 6 and contact 9.

directed parallel to their magnetic easy axis. The sizes of Py1 and Py2 in batch I were $2.0 \times 0.8\text{ }\mu\text{m}^2$ and $14 \times 0.5\text{ }\mu\text{m}^2$ respectively, as shown in Fig. 1a. An additional set of Py1 and Py2, with sizes of $2.0 \times 0.5\text{ }\mu\text{m}^2$ and $14 \times 0.1\text{ }\mu\text{m}^2$, was used in batch II. This set showed an improved magnetic switching behaviour and had coercive fields three times larger.

In the second fabrication step, 50-nm -thick crossed Cu strips were deposited by e-gun evaporation in 1×10^{-8} mbar vacuum. Before Cu deposition, the oxide of the Py electrodes was removed by ion milling, to ensure transparent contacts. The conductivities of the Py and Cu films were determined to be $\sigma_{\text{Py}} = 6.6 \times 10^6\text{ }\Omega^{-1}\text{ m}^{-1}$ and $\sigma_{\text{Cu}} = 3.5 \times 10^7\text{ }\Omega^{-1}\text{ m}^{-1}$ at room temperature. At 4.2 K both conductivities increased by a factor of 2.

The measurements were performed by standard a.c. lock-in techniques, using current magnitudes of $100\text{ }\mu\text{A}$ to 1 mA . We note that in the 'conventional' spin-valve geometry (sending a current I from contacts to 1 to 7, and measuring the voltage V between contacts 4 and 9), the signal $R = V/I$ was completely dominated by AMR and Hall effects of the Py contacts, having a typical magnitude of $10\text{ m}\Omega$. Although we have used the AMR signal of the contacts to confirm the switching of the Py electrodes, its presence has made

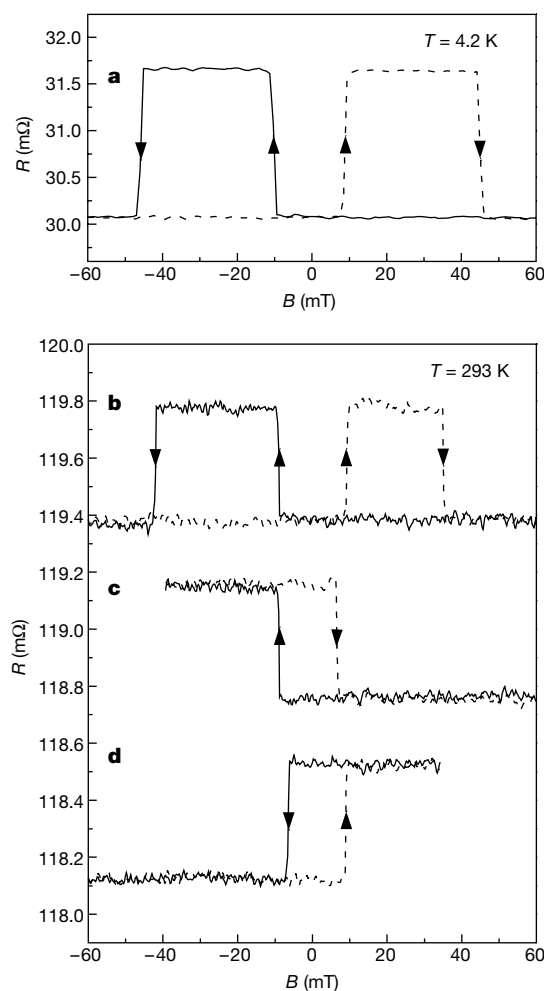


Figure 2 The spin valve effect at $T = 4.2\text{ K}$ (**a**) and room temperature (**b**) in the non-local geometry for a sample with 250 nm Py electrode spacing. An increase in resistance is observed, when the magnetization configuration is changed from parallel to anti-parallel. The solid (dashed) lines correspond to the negative (positive) sweep direction. **c**, **d**, The 'memory effect'. For clarity **c** and **d** are offset downwards. We note that the vertical scale of **a** is different from that of **b**, **c** and **d**. The sizes of the Py1 and Py2 electrodes are $2.0 \times 0.5\text{ }\mu\text{m}^2$ and $14 \times 0.1\text{ }\mu\text{m}^2$.

any observation of a spin-valve signal in the conventional spin-valve geometry impossible.

However, in the non-local spin-valve geometry (see Fig. 1b) these spurious effects can be eliminated. This technique is similar to the “potentiometric” method of Johnson^{13,14}. However, in our case, we have completely separated the current and the voltage circuits. Current now enters from contact 1 (Py1) and is extracted at contact 5 (Cu), whereas the voltage is measured between contact 6 (V_{Cu}) and contact 9 (V_{Py2}). Changing the relative magnetization configuration of Py1 and Py2 from parallel to anti-parallel will yield a signal only if the densities (electrochemical potentials) of the spin-up and spin-down electrons in the centre of the Cu cross are unequal. In a parallel configuration Py2 will measure the highest value of the two unequal spin electrochemical potentials, whereas in an anti-parallel configuration it will measure the lower value. Changing from a parallel to an anti-parallel configuration will therefore decrease the voltage (V_{Py2}) measured by Py2, resulting in an increase in the voltage difference ($V_{Cu} - V_{Py2}$) measured, and hence an increase of the resistance.

Figure 2a and b shows typical data taken at 4.2 K and room temperature for a sample from batch II, with a 250-nm Py electrode spacing. While sweeping the magnetic field from negative to positive field, we observed an increase in the resistance when the magnetization of Py1 flips at 9 mT, resulting in an anti-parallel magnetization configuration. When the magnetization of Py2 flips at 47 mT (4.2 K) and 38 mT (room temperature), the magnetizations are parallel again, but now point in the opposite direction. The magnitude of the measured background resistance, around 30 m Ω at 4.2 K and 120 m Ω at room temperature, depends on the geometrical shape of the Cu cross and is typically a fraction of the Cu square resistance.

Figure 2c and d shows the ‘memory effect’. Coming from high positive **B** field, the sweep direction of the **B** field is reversed after Py1 has switched, but Py2 has not. At the moment of reversing the sweep direction, the magnetic configuration of Py1 and Py2 is anti-parallel, and accordingly a higher resistance is measured. When the **B** field is swept back to its original high positive value, the resistance remains at its increased level until Py1 switches back at a positive field of 9 mT. At zero **B** field the resistance can therefore have two distinct values, depending on the magnetic history of the Py electrodes. Samples with larger Py electrode spacing show identical switching behaviour, but the magnitude of the spin signal ΔR is reduced, as we will discuss below.

We have calculated the theoretically expected magnitude and the Py electrode distance dependence of the spin valve signal ΔR for the non-local geometry of Fig. 1b in the diffusive transport regime and

for transparent interfaces, following the lines of the standard Valet Fert model for GMR^{20,21}, adopted for our multi-terminal geometry. We obtain²²:

$$\Delta R = \frac{\alpha_F^2 \frac{\lambda_N}{\sigma_N A} e^{(-L/2\lambda_N)}}{(M+1)[M \sinh(L/2\lambda_N) + \cosh(L/2\lambda_N)]} \quad (1)$$

where $M = (\lambda_N \sigma_F / \lambda_F \sigma_N)(1 - \alpha_F^2)$, $\alpha_F = \sigma_\uparrow - \sigma_\downarrow / \sigma_\uparrow + \sigma_\downarrow$ is the bulk current polarization of the Py electrodes, $\sigma_\uparrow$ ($\sigma_\downarrow$) are spin up (down) conductivities in the ferromagnet, σ_N (σ_F) is the total conductivity of the normal metal (ferromagnetic metal), λ_N (λ_F) is the spin-flip length in the normal metal (ferromagnetic metal), L is the distance between the two Py electrodes, and A is the cross-sectional area of the normal metal wire. Equation (1) shows that for $\lambda_N \ll L$, the magnitude of the spin signal ΔR will decay exponentially as a function of L . In the opposite limit, $\lambda_F \ll L \ll \lambda_N$, the spin signal ΔR has a $1/L$ dependence. We note that the spin signal ΔR is determined by the bulk conductances of the ferromagnet and normal metal over a distance of the spin-flip lengths. The origin of the spin signal is therefore different from the spin-dependent transport in tunnel junction experiments, where the density of states at the interface is the important quantity for the magnitude of the spin signal²³. Hence, for tunnel junctions, the nature of the first atomic layers near the interface is crucial, whereas for transparent contacts, which is the case for our device, the interface properties are expected to be less important^{15,21}.

We have measured the reduction of the magnitude of spin signal ΔR as a function of the Py electrode spacing L , as shown in Fig. 3. By fitting the data to equation (1) we have obtained λ_N in the Cu wire. From the best fits we find $\lambda_N = 1 \pm 0.2 \mu\text{m}$ at $T = 4.2 \text{ K}$, and $\lambda_N = 350 \pm 50 \text{ nm}$ at room temperature. The values of λ_N are compatible with those reported in the literature²⁴: $\lambda_N \approx 450 \text{ nm}$ for Cu in GMR measurements at 4.2 K. However, we cannot make a straightforward comparison between the GMR results and ours. In the thin films we use, the elastic mean free path of the electrons is limited by surface scattering, causing the conductivity of the Cu to be smaller than in GMR layers. We also note that in Fig. 3 no good fit can be obtained for data where $L = 250 \text{ nm}$. As the Py electrode spacing L approaches the width W of the Cu wire, the presence of the side arms (Fig. 1a) will give rise to a local enhancement of the conductance at the Cu cross and hence can result in an increase in the spin signal. Therefore, in this limit, we can expect deviations from our one-dimensional model.

We can calculate the spin-flip times τ_{sf} in the Cu wire, using a Fermi velocity of $v_F = 1.57 \times 10^6 \text{ m s}^{-1}$ (ref. 25). At 4.2 K we find $\tau_{sf} = 42 \text{ ps}$, while at room temperature $\tau_{sf} = 11 \text{ ps}$. We will not discuss the physics of the spin-flip scattering process here. However, comparing the spin-flip time to the elastic scattering time $\tau_e = 2.9 \times 10^{-14} \text{ s}$ at 4.2 K, we find that on average the spin is flipped after about 10^3 elastic scattering events in the Cu wire.

In principle the fits of Fig. 3 also yield the spin polarization α_F and the spin-flip length λ_F of the Py electrodes. However, the values of α_F and λ_F cannot be determined separately, because in the relevant limit ($M \gg 1$) which applies to our experiment ($12 \leq M \leq 26$), the spin signal ΔR is proportional to the product $\alpha_F^2 \lambda_F^2$. From the fits we find $\alpha_F \lambda_F \approx 1.2 \text{ nm}$ at 4.2 K and $\alpha_F \lambda_F \approx 0.5 \text{ nm}$ at room temperature. Taking, from refs 17 and 18, a spin-flip length in the Py electrode of $\lambda_F = 5.5 \text{ nm}$ (at 4.2 K), a bulk current polarization of 22% in the Py electrodes at 4.2 K is obtained: $\alpha_F \approx 0.22$. These values are in the same range as the results obtained from the analysis of the GMR effect^{9,10,17,18,24}.

With the obtained parameters we can calculate the maximum current polarization $P = I_\uparrow - I_\downarrow / I_\uparrow + I_\downarrow$ in the Cu wire. For the samples with the smallest Py electrode spacing we obtain $P \approx 2\%$ at 4.2 K. When we scale the observed signals to the device cross-sections, we find that the scaled spin-valve signal of refs 13 and 14 is typically four orders of magnitude larger than ours. This contrast

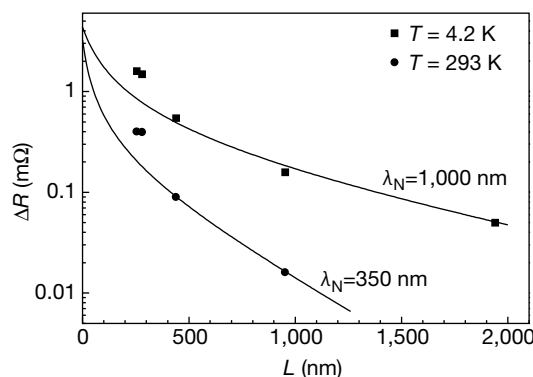


Figure 3 Dependence of the magnitude of the spin signal ΔR on the Py electrode distance L . The solid squares represent data taken at $T = 4.2 \text{ K}$; the solid circles represent data taken at room temperature. The solid lines represent the best fits based on equation (1).

corresponds with the need to invoke a spin polarization of the current in the normal metal of about 100%, to explain the results of refs 13 and 14 in terms of spin accumulation. In our opinion this must imply that the observed effects of refs 13 and 14 cannot be related to spin accumulation.

We have thus demonstrated spin injection and accumulation in a mesoscopic spin valve. We find a surprisingly long spin-flip length in Cu of around 1 μm at 4.2 K and about 350 nm at room temperature. For the smallest Py electrode spacing, the magnitude of the spin signal and the current polarization P in the Cu wire are limited by the unfavourable ratio of the spin-independent resistance of the Cu strips (L/σ_N) and the spin-dependent resistance of the Py ferromagnet (λ_F/σ_F). In principle, larger signals can therefore be obtained by a proper choice of materials and geometries.

Finally we note that our system permits the study of spin transport phenomena, such as controlled spin precession in solid state devices and the control of spin-polarized currents at room temperature by additional ferromagnetic contacts^{3,26}. □

Received 3 November 2000; accepted 2 January 2001.

1. Prinz, G. A. Spin-polarized transport. *Phys. Today* **48**, 58–63 (1995).
2. Prinz, G. A. Magnetoelectronics. *Science* **282**, 1660–1663 (1998).
3. Brataas, A., Nazarov, Y. N. & Bauer, G. E. W. Finite-element theory of transport in ferromagnet-normal metal systems. *Phys. Rev. Lett.* **84**, 2481–2484 (2000).
4. Johnson, M. & Silsbee, R. H. Interfacial charge-spin coupling: injection and detection of spin magnetization in metals. *Phys. Rev. Lett.* **55**, 1790–1793 (1985).
5. Kikkawa, J. M. & Awschalom, D. D. Lateral drag of spin coherence in gallium arsenide. *Nature* **397**, 139–141 (1999).
6. Fiederling, R. *et al.* Injection and detection of a spin-polarized current in a light-emitting diode. *Nature* **402**, 787–790 (2000).
7. Ohno, Y. *et al.* Electrical spin injection in a ferromagnetic semiconductor heterostructure. *Nature* **402**, 790–792 (2000).
8. Malajovich, I., Kikkawa, J. M., Awschalom, D. D. Coherent transfer of spin through a semiconductor heterointerface. *Phys. Rev. Lett.* **84**, 1015–1018 (2000).
9. Gijs, M. A. M. & Bauer, G. E. W. Perpendicular giant magnetoresistance of magnetic multilayers. *Adv. Phys.* **46**, 285–445 (1997).
10. Ansermet, J.-Ph. Perpendicular transport of spin-polarized electrons through magnetic nanostructures. *J. Phys. Cond. Matter* **10**, 6027–6050 (1998).
11. Ebels, U., Radulescu, A., Henry, Y., Pirau, L. & Ounadjela, K. Spin accumulation and domain wall magnetoresistance in 35 & nm Co wires. *Phys. Rev. Lett.* **84**, 983–986 (2000).
12. Katine, J. A., Albert, F. J., Buhrmann, R. A., Myers, E. B. & Ralph, D. C. Current driven magnetization reversal and spin wave excitations in Co/Cu/Co pillars. *Phys. Rev. Lett.* **84**, 3149–3152 (2000).
13. Johnson, M. Spin accumulation in gold films. *Phys. Rev. Lett.* **70**, 2142–2145 (1993).
14. Johnson, M. Bipolar spin switch. *Science* **260**, 320–322 (1993).
15. Fert, A. & Lee, S. Theory of the bipolar spin switch. *Phys. Rev. B* **53**, 6554–6565 (1996).
16. Herschfield, S., Zhao, L. Z. Charge and spin transport through a metallic ferromagnetic-paramagnetic-ferromagnetic junction. *Phys. Rev. B* **56**, 3296–3305 (1997).
17. Steenwyk, S. D., Hsu, S. Y., Loloe, R., Bass, J. & Pratt Jr, W. P. Perpendicular-current exchanged-biased spin-valve evidence for a short spin-diffusion length in permalloy. *J. Magn. Magn. Mater.* **170**, L1–L6 (1997).
18. Dubois, S. *et al.* Evidence for a short spin diffusion length in permalloy from the giant magnetoresistance of multilayered nanowires. *Phys. Rev. B* **60**, 477–484 (1999).
19. Crangle, J. *The Magnetic Properties of Solids* Ch. 6 (Edward Arnold, London, 1977).
20. Van Son, P., van Kempen, H. & Wyder, P. Boundary resistance of the ferromagnetic-nonferromagnetic metal interface. *Phys. Rev. Lett.* **58**, 2271–2273 (1987).
21. Valet, T. & Fert, A. Theory of the perpendicular magnetoresistance in magnetic multilayers. *Phys. Rev. B* **48**, 7099–7113 (1993).
22. Filip, A. T., Hoving, B. H., Jedema, F. J. & Van Wees, B. J. Experimental search for the electrical spin injection in a semiconductor. *Phys. Rev. B* **62**, 9996–9999 (2000).
23. Meservey, R. & Tedrow, P. M. Spin-polarized electron tunneling. *Phys. Rep.* **238**, 173–243 (1994).
24. Yang, Q. *et al.* Spin flip diffusion length and giant magnetoresistance at low temperatures. *Phys. Rev. Lett.* **72**, 3274–3277 (1994).
25. Ahscroft, N. W. & Mermin, N. D. in *Solid State Physics* Ch. 2 (ed. Crane, D. G.) 38 (W. B. Saunders, Orlando, 1976).
26. Hernando, D. H., Nazarov, Yu. V., Brataas, A. & Bauer, G. E. W. Conductance modulation by spin precession in noncollinear ferromagnetic normal-metal ferromagnetic systems. *Phys. Rev. B* **62**, 5700–5712 (2000).

Acknowledgements

We thank H. Boeve, J. Das and J. de Boeck at IMEC (Belgium) for support in sample fabrication, M. S. Nijboer for experimental assistance and T. M. Klapwijk for discussions. We wish to thank the Stichting Fundamenteel Onderzoek der Materie and the EU project SPIDER for financial support.

Correspondence should be addressed to F.J.J. (e-mail: jedema@phys.rug.nl).

Shear-induced molecular precession in a hexatic Langmuir monolayer

Jordi Ignés-Mullol* & Daniel K. Schwartz*

Department of Chemistry, Tulane University, New Orleans, Louisiana 70118, USA

* Present addresses: Department de Química Física, Universitat de Barcelona, Martí i Franquès I, Barcelona E08028, Spain (J.I.-M.); Department of Chemical Engineering, University of Colorado, Boulder, Colorado 80309, USA (D.K.S.)

Liquid crystalline behaviour is generally limited to a select group of specially designed bulk substances. By contrast, it is a common feature of simple molecular monolayers and other quasi-two-dimensional systems¹, which often possess a type of in-plane ordering that results from unbinding of dislocations—a ‘hexatic’ liquid crystalline phase. The flow of monolayers is closely related to molecular transport in biological membranes, affects foam and emulsion stability and is relevant to microfluidics research. For liquid crystalline phases, it is important to understand the coupling of the molecular orientation to the flow. Orientationally ordered (nematic) phases in bulk liquid crystals exhibit ‘shear aligning’ or ‘tumbling’ behaviour under shear, and are described quantitatively by Leslie–Ericksen theory². For hexatic monolayers, the effects of flow have been inferred from textures of Langmuir–Blodgett films^{3–5} and directly observed at the macroscopic level^{6–10}. However, there is no accepted model of hexatic flow at the molecular level. Here we report observations of a hexatic Langmuir monolayer that reveal continuous, shear-induced molecular precession, interrupted by occasional jump discontinuities. Although superficially similar to tumbling in a bulk nematic phase, the kinematic details are quite different and provide a possible mechanism for domain coarsening and eventual molecular alignment in monolayers. We explain the precession and jumps within a quantitative framework that involves coupling of molecular orientation to the local molecular hexatic ‘lattice’, which is continuously deformed by shear.

The interfacial thermodynamics and rheology of molecular films like Langmuir monolayers are responsible for the stability of multiphase materials such as foams and emulsions¹¹. They are important model systems for studies of dynamics and molecular interactions in bio-membranes and, as quasi-two-dimensional systems, display a variety of unusual liquid crystalline phenomena. The application of synchrotron X-ray scattering and specialized imaging techniques such as fluorescence and Brewster angle microscopy (BAM) have

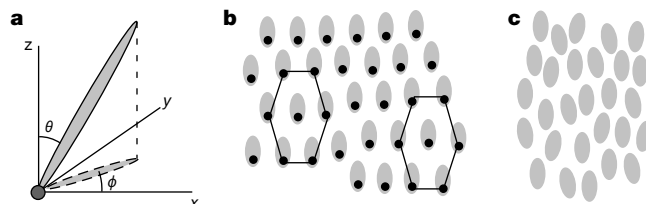


Figure 1 Hexatic and nematic phase representations. **a**, A cartoon showing the molecular geometry of rod-shaped molecules in a tilted hexatic phase. The molecule is tilted away from the surface normal by the angle θ , and the azimuthal direction of tilt is given by the angle ϕ . **b**, Schematic picture of a tilted hexatic phase viewed from above. The projection of the molecules on the surface plane are oval; the dark circles represent the headgroups at the interface. In this case, the molecular tilt direction is locked into the nearest neighbour direction of the local distorted-hexagonal unit cell. We note that the orientation of the unit cell remains uniform despite the presence of a lattice dislocation that destroys the positional registry of the lattice. **c**, Schematic picture of a nematic phase. Molecules are aligned in a common direction on average, but their positions are random.

greatly improved our understanding of monolayer structure and phase behaviour in the last decade¹². Hexatic phases were found to be extremely common in monolayers and, in fact, characteristic of these systems. However, a variety of recently discovered dynamic and rheological phenomena, such as the coupling of molecular orientation to flow^{6–10,13,14}, unusual relaxation dynamics and shear moduli^{15–17}, and non-newtonian velocity profiles in channel flow^{18–20}, remain largely unexplained. Another mysterious effect of flow is that it may have either an annealing or a fragmenting effect on the monolayer domain structure (L'_2 and Ov phases respectively^{6,7}).

Many hexatic monolayer phases are characterized by a finite molecular tilt at a polar angle, θ , with respect to the surface normal (see Fig. 1a). A typical monolayer sample consists of a mosaic of macroscopic domains separated by sharp boundaries; molecules within a given domain share a common azimuthal tilt direction, φ (Fig. 1a). In these phases, the tilt direction is directly linked to the orientation of the local pseudo-hexagonal unit cell (Fig. 1b). The orientation of the unit cell extends over macroscopic distances, although true crystalline order is destroyed by the presence of local point defects (lattice dislocations). The molecules in a nematic phase (Fig. 1c), on the other hand, are aligned on average in a particular direction; however, their positions are completely disordered. The coupling of molecular orientation to flow is well understood for bulk nematic phases², and is important in the processing of many polymeric and/or fibre-based materials. There is evidence that the description of nematic flow can be extended to two dimensions. Workers^{21,22} who studied the transient behaviour of a nematic monolayer (of a hairy-rod polymer) under extensional flow found good agreement with a two-dimensional rigid rod model. There is no general understanding or accepted picture of the coupling of structure to flow in hexatic phases.

Because the intensity (grey level) within a BAM image is directly related to φ , BAM provides a convenient method of studying changes in the molecular orientation. Previous observations in fatty acid monolayers suggested that flow induced a reorientation of the molecules to a particular angle^{6–8} with respect to the flow. The mechanism for this “alignment” was the sudden appearance of new domains with the appropriate orientation. The domains grew by movement of a boundary; the propagating fronts were described as shear bands^{6,7}. We report a qualitatively different phenomenon: a continuous change in the reflectivity of monolayer domains corresponding to a cooperative, continuous rotation of the molecular orientation.

Figure 2 shows a time sequence of BAM images corresponding to a simple shear flow experiment⁸ on a docosanoic acid

($\text{CH}_3(\text{CH}_2)_{20}\text{COOH}$) monolayer in the L'_2 phase¹² ($T = 17^\circ\text{C}$). Before the onset of flow (Fig. 2a), the central fish-shaped domain is darker than the surrounding region. As shear is initiated (Fig. 2a, b) the shape of the domain is gradually skewed and the contrast is reduced between the domain and the surroundings. At some point the contrast is almost completely lost (Fig. 2d). As shear continues, the original domain reappears (with a distorted shape), but the contrast has been inverted. Qualitatively, this sequence corresponds to a simultaneous anticlockwise rotation of the molecules within the central domain from left to right and of the molecules in the surrounding region from right to left. This particular example of contrast inversion is clear evidence of continuous molecular precession. However, neighbouring domains do not generally rotate simultaneously, or in phase with each other. The independence of domain dynamics suggests that the precession is not related to domain–domain interactions. Our observation, in recent experiments, that domains can ‘slip’ along each other during shear is also consistent with a picture where the molecular details of neighbouring domains are de-coupled.

We have measured the evolution of the grey level during 35 precession events. In all cases, after what can be long periods of time in an initial state where the tilt azimuth is approximately parallel to flow, the grey level of a domain changes relatively quickly from dark to light (corresponding to a change in molecular orientation from the left to the right of the incidence plane), or vice versa. This corresponds to a nearly 180° rotation of the molecules, after which the grey level remains essentially constant again. Representative data presented in Fig. 3 show that a higher shear rate ($\dot{\gamma}$) results in faster molecular precession (Fig. 3a). In fact, the rate of change of the reflectivity is proportional to the shear rate (see Fig. 3b) and the data can be collapsed when plotted as a function of the strain ($\gamma = \dot{\gamma}t$). This indicates a flow-induced reorientation of the alkane tails. The scaling with shear rate also suggests that our observations are made under steady state conditions, where elastic effects can be ignored. The qualitative behaviour, that is, fast change in reflectivity followed by saturation (more clearly observable in Fig. 4) is reminiscent of a Jeffery orbit²³, which describes the rotation of an elongated object under shear or the tumbling of nematic molecules. The Jeffery orbit is periodic and predicts that the domain will eventually return back to its original intensity. A quantitative comparison, however, reveals that the observed dynamics cannot be reconciled with a Jeffery orbit. In particular, a Jeffery orbit that is consistent with the steepest region of the curves in Fig. 3 predicts that the intensity will pass a maximum and start to decrease well within the long period of time during which the observed intensity remains constant. Our observation of asymptotic regimes before and after rapid change instead

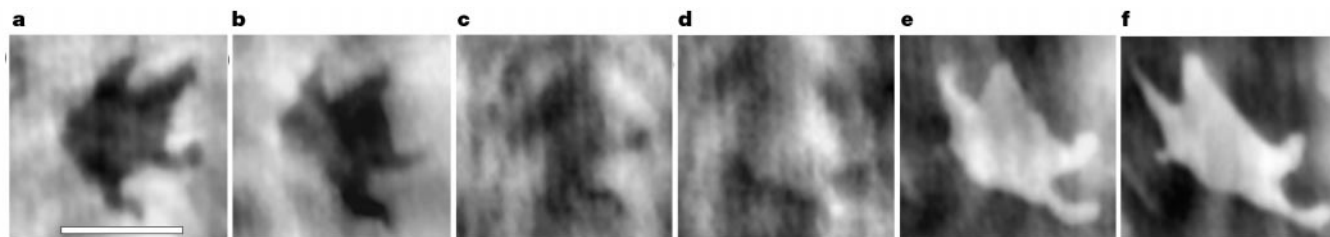


Figure 2 Continuous change in the reflectivity of domains in a docosanoic acid ($\text{CH}_3(\text{CH}_2)_{20}\text{COOH}$) monolayer in the L'_2 phase. ($\Gamma = 23 \text{ mN m}^{-1}$). Shear is counter-clockwise ($\dot{\gamma} = -0.2 \text{ s}^{-1}$) and the flow field is horizontal. Images are obtained with a BAM with the laser incident from the top (perpendicular to the flow) and the analyser set at 60° for optimal contrast. Scale bar in **a** is $100 \mu\text{m}$ long. The grey level of the images is linearly related to the local reflectivity. Qualitatively, the molecular azimuth in the dark domains is oriented towards the left of the incidence plane and the light domains towards the right. Before this sequence, the monolayer had been prepared through a series of shear reversal cycles (in which the flow is inverted after half a cycle), resulting in the

reduction of the initially large number of reflectivities in the mosaic of domains to only two different reflectivities. The dark domain in the centre of **a** exhibits a continuous increase of its reflectivity, while the surrounding domain darkens as strain increases corresponding to continuous anticlockwise precession of the molecular orientation in both domains. Elapsed times after frame **a** are as follows: **b**, $t = 1.7 \text{ s}$; **c**, $t = 2.9 \text{ s}$; **d**, $t = 4.3 \text{ s}$; **e**, $t = 5.4 \text{ s}$; and **f**, $t = 6.8 \text{ s}$. The contrast drops to zero between frames **c** and **d** as the reflectivities of both domains become equal. Optical artefacts such as interference fringes and illumination gradients are present and unavoidable.

of extrema followed by periodic behaviour indicates a clear incompatibility with the nematic tumbling model.

Our experiments also reveal the presence of jump discontinuities in the otherwise continuous change in the reflectivity of the monolayer (Fig. 4). They appear as a diffuse front that propagates in arbitrary directions, changing the reflectivity of the affected domain (this is clearly different for the sharp and well-oriented alignment fronts observed in other monolayer phases). In many instances, the magnitude of these jumps is close to 30° , as discussed below.

Both the presence of discontinuous orientational jumps and the quantitative details of the precession dynamics are inconsistent with expected behaviour for a tumbling nematic phase. Although we can imagine the projection of a tilted molecule onto the surface plane as an analogue of the nematic director, the structural details of hexatic monolayer phases are quite different from those of a nematic. In particular, there is a strong coupling between the tilt azimuth and the orientation of the local pseudo-hexagonal unit cell, the hexatic bond-orientational order parameter^{12,24,25}. Both L'_2 and O_v phases have molecules tilted towards their next nearest neighbour (NNN) while the L_2 phase is characterized by a tilt towards a nearest neighbour (NN) molecule. An attempt to describe the dynamics of these phases using a nematic model completely neglects the existence of hexatic order and the coupling of the molecular tilt to the local lattice.

Although shear flow continuously distorts the shape of the underlying lattice, we expect that the tilt azimuth of a given molecule will prefer to remain in a position that approximates the NNN environment. This is consistent with our observation that cessation of the flow at any time during shear gives a configuration whose domain reflectivities do not appear to relax with time. As a first-order model, we propose that the tilt azimuth, φ , of a given molecule attempts to follow the displacement of its NNN as the lattice is

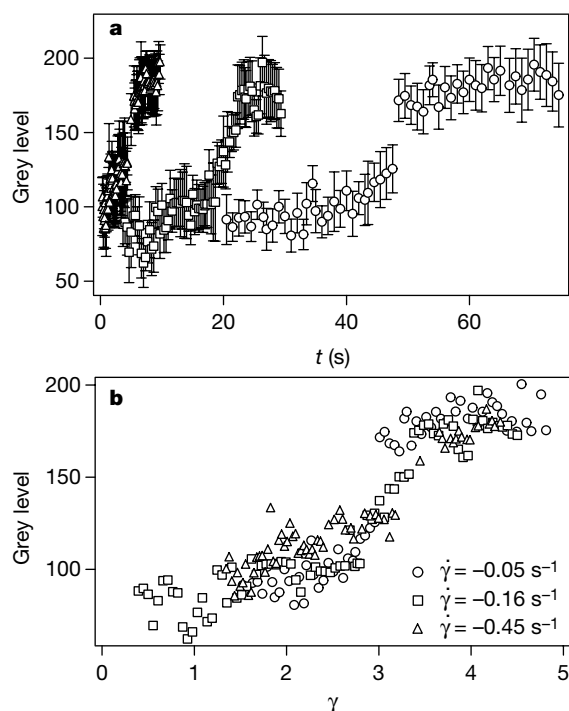


Figure 3 Evolution of the 8-bit digitized reflectivity for three different domains under three different shear flows. The monolayers are in the L'_2 phase. ($\Pi = 20 \text{ mN m}^{-1}$). **a**, The mean (data point) and the standard deviation (error bar) of the instantaneous distribution of grey levels of the domain under observation are shown combined. **b**, The same data are shown after the time axis is transformed into the strain axis. The data sets are superimposed by adding arbitrary offsets to the strain axis.

deformed by shear (Fig. 5). This leads to the relationship $\cotan[\varphi(t)] = \dot{\gamma}(t - t_0)$, with $t_0 = -x_0/(\dot{\gamma}y_0)$. This expression is qualitatively consistent with our observation of a steep change in the reflectivity of the domains preceded and followed by asymptotic regimes (Figs 2 and 3). This continuous model, however, is topologically inconsistent with the NNN lattice-bond symmetry condition. Indeed, as the lattice is distorted by shear, the relative position of different sites will change and sites that originally were NNN actually become NN (see Fig. 5); not a stable tilt direction in the L'_2 phase. The observed jump discontinuities in the evolution of $\varphi(t)$ resolve this inconsistency. Composing the expression for $\varphi(t)$ with the calculated reflectivity as a function of φ , we obtain an expression that is fitted to the data in Figs 3 and 4. In this way, we

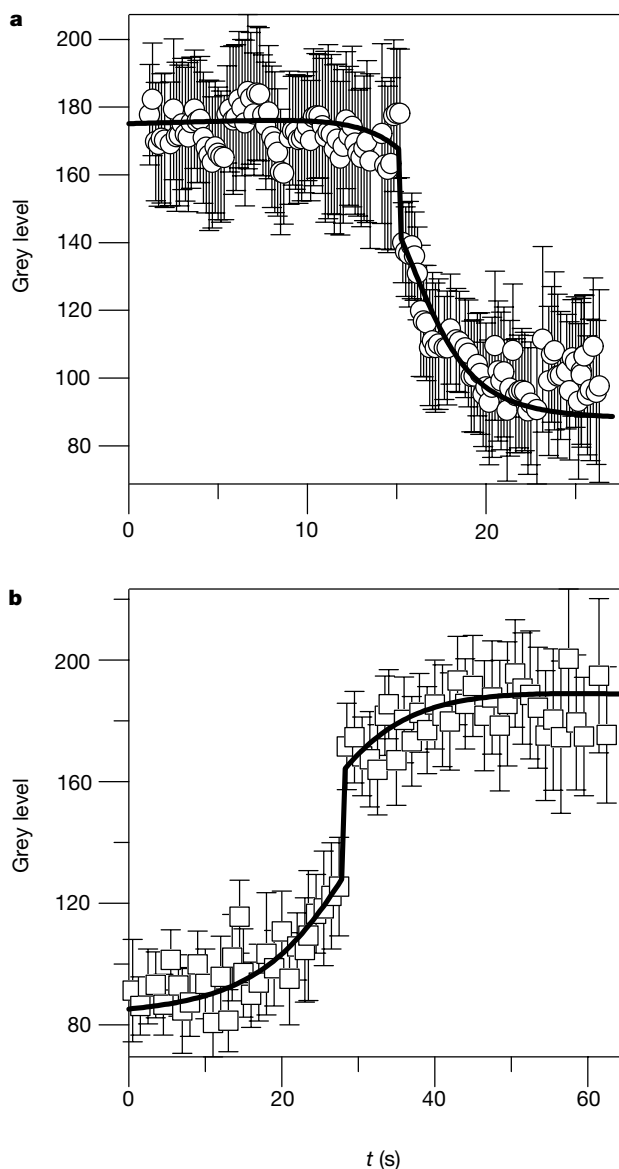


Figure 4 Jump discontinuities are revealed in the reflectivity of two domains from different experiments shearing a monolayer in the L'_2 phase. **a**, $\Pi = 22 \text{ mN m}^{-1}$, $\dot{\gamma} = 0.2 \text{ s}^{-1}$ (clockwise); **b**, $\Pi = 20 \text{ mN m}^{-1}$, $\dot{\gamma} = -0.05 \text{ s}^{-1}$ (anticlockwise). The solid lines are fits to the data using the model proposed in the text and that characterize the jumps as follows: **a**, $\varphi = 311^\circ \pm 7^\circ$ to $\varphi = 278^\circ \pm 7^\circ$ ($\Delta\varphi = -33^\circ \pm 10^\circ$); **b**, $\varphi = 262^\circ \pm 6^\circ$ to $\varphi = 294^\circ \pm 6^\circ$ ($\Delta\varphi = 32^\circ \pm 9^\circ$). Our 35 observed precession events have clearly revealed 23 instances of jump discontinuities. We have obtained a quantitative estimation for the magnitude of the jump in 17 of those cases, which resulted in 11 jumps of magnitude close to 30° .

estimate the magnitude of the experimentally observed jump discontinuities, many of which are close to 30° .

Figure 5 provides a geometrical interpretation for the observed magnitude of the largest jumps, which we are most likely to observe. The jump depicted in Fig. 5 has a magnitude of 26° within an angular range where the variation in BAM intensity is large, consistent with the majority of our observations. This picture also predicts that a number of smaller jump discontinuities ($<15^\circ$) will continue to occur as large strains are applied in order to reconcile the NNN lattice-bond orientation and the lattice distortion. These jumps would take place in a region of reduced reflectivity and would be smaller in magnitude, making them more difficult to observe.

We believe that the macroscopic appearance of monolayers following flow is directly related to the coupling between flow and molecular orientation. The combination of continuous precession and jump discontinuities of molecular orientation in the L_2' phase allows gradual alignment of molecular orientation in adjacent domains, resulting, ultimately, in a coarsened domain structure with two predominant molecular orientations. However, the

nucleation and propagation of new domains upon initiation (or reversal) of flow in the Ov phase breaks up previously coherent domains, leading to increasing fragmentation.

The details of the molecular precession reported here combined with our previous observations regarding shear alignment of the molecular lattice in the Ov phase⁸ suggest a conceptual framework for the flow of hexatic phases that is fundamentally different from that for nematic phases. Although qualitatively similar phenomena are observed, that is, shear-alignment and continuous orientation rotation, they are distinctly different when analysed quantitatively. Flow acts directly on the molecular orientation (director) in a nematic phase whereas the local unit cell orientation dominates the behaviour in a hexatic monolayer; the molecular azimuth is essentially locked into the lattice. This concept can be used to guide our thinking about the effects of flow in quasi-two-dimensional systems (monolayers, biological membranes, and so on), in which hexatic phases are so prevalent. □

Received 13 November 2000; accepted 9 January 2001.

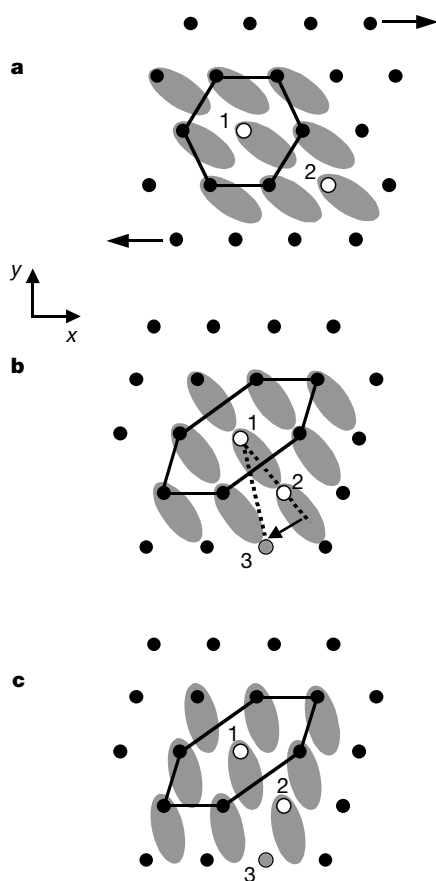


Figure 5 Geometrical model for the onset and magnitude of the jump discontinuities in the evolution of $\varphi(t)$. This picture assumes that the degenerate lattice lines align with the flow. This is the same lattice alignment that was observed as a result of shear alignment in the Ov phase⁸ and is consistent with the observations reported here. The flow field (horizontal clockwise shear) relative to the reference site (1) is illustrated by the arrows in **a**. **a**, A region that initially satisfies the next nearest neighbour (NNN) lattice-bond orientation is considered. The azimuth in the reference site (1) is towards a NNN (2), whose initial coordinates with respect to (1) are (x_0, y_0) . These coordinates change to $(x_0 + \gamma t y_0, y_0)$ when a strain γt has been applied. **b**, After a strain of $\gamma = 0.6$ the continuous reorientation results in a violation of the NNN condition, as (1) and (2) now become NN. **c**, A jump in $\varphi(t)$ can result in a more favourable configuration, closer to the NNN, with (1) tilted towards (3). In this example, the jump takes place between $\varphi = 310^\circ$ and $\varphi = 284^\circ$ ($\Delta\varphi = -26^\circ$), consistent with the data in Fig. 4a.

- Halperin, B. I. & Nelson, D. R. Theory of two-dimensional melting. *Phys. Rev. Lett.* **41**, 121–124 (1978).
- Chandrasekhar, S. *Liquid Crystals* 2nd edn (Cambridge Univ. Press, Cambridge/New York, 1992).
- Peterson, I. R. Optical observation of monomer Langmuir-Blodgett film structure. *Thin Solid Films* **116**, 357–365 (1984).
- Leuthe, A. & Riegler, H. Thermal behaviour of Langmuir-Blodgett films. II. X-ray and polarized reflection microscopy studies on coexisting polymorphism, thermal annealing and epitaxial layer growth of behenic acid multilayers. *J. Phys. D* **25**, 1786–1797 (1992).
- Bubeck, C. Imaging of the lateral structure of Langmuir-Blodgett films by metal decoration and polarization microscopy. *Thin Solid Films* **210/211**, 674–677 (1992).
- Maruyama, T., Fuller, G., Frank, C. & Robertson, C. Flow-induced molecular orientation of a Langmuir film. *Science* **274**, 233–235 (1996).
- Maruyama, T., Lauger, J., Fuller, G. G., Frank, C. W. & Robertson, C. R. Orientation in a fatty acid monolayer: effect of flow type. *Langmuir* **14**, 1836–1845 (1998).
- Ignés-Mullol, J. & Schwartz, D. K. Alignment of hexatic Langmuir monolayers under shear. *Phys. Rev. Lett.* **85**, 1476–1479 (2000).
- Ikegami, K., Mingotaud, C. & Delhaes, P. Monolayer flow and in-plane orientation induced by a rotating disk in Langmuir and Langmuir-Blodgett films of a merocyanine dye. *Phys. Rev. E* **56**, 1987–1997 (1997).
- Ikegami, K., Mingotaud, C. & Delhaes, P. In-plane orientation of mesoscopic domains in Langmuir and Langmuir-Blodgett films induced by rotating disks. *Thin Solid Films* **329**, 24–27 (1998).
- Edwards, D. A., Brenner, H. & Wasan, D. T. *Interfacial Transport Processes and Rheology* (Butterworth-Heinemann, Boston, 1991).
- Kaganer, V. M., Möhwald, H. & Dutta, P. Structure and phase transitions in Langmuir monolayers. *Rev. Mod. Phys.* **71**, 779–819 (1999).
- Friedenberg, M. C., Fuller, G. G., Frank, C. W. & Robertson, C. R. Direct visualization of flow-induced anisotropy in a fatty acid monolayer. *Langmuir* **12**, 1594–1599 (1996).
- Cuvillier, N., Mingotaud, C. & Ikegami, K. Shear-induced optical anisotropy in a Langmuir monolayer: A Brewster angle reflectivity study. *J. Chem. Phys.* **111**, 6982–6990 (1999).
- Ghaskadvi, R. S., Bohanon, T. M., Dutta, P. & Ketterson, J. B. Shear response of Langmuir monolayers of heneicosanoic (C-21) acid studied using a torsion pendulum. *Phys. Rev. E* **54**, 1770–1773 (1996).
- Ghaskadvi, R. S., Ketterson, J. B. & Dutta, P. Nonlinear shear response and anomalous pressure dependence of viscosity in a Langmuir monolayer. *Langmuir* **13**, 5137–5140 (1997).
- Ghaskadvi, R. S., Carr, S. & Dennin, M. Effect of subphase Ca^{++} ions on the viscoelastic properties of Langmuir monolayers. *J. Chem. Phys.* **111**, 3675–3678 (1999).
- Kurnaz, M. L. & Schwartz, D. K. Channel flow in a Langmuir monolayer: unusual velocity profiles in a liquid-crystalline mesophase. *Phys. Rev. E* **56**, 3378–3384 (1997).
- Ivanova, A., Kurnaz, M. L. & Schwartz, D. K. Temperature and flow rate dependence of the velocity profile during channel flow of a Langmuir monolayer. *Langmuir* **15**, 4622–4624 (1999).
- Ivanova, A. & Schwartz, D. K. Transient behavior of the velocity profile in channel flow of a Langmuir monolayer. *Langmuir* **16**, 9433–9438 (2000).
- Maffettone, P. L., Grosso, M., Friedenberg, M. C. & Fuller, G. G. Extensional flow of a two-dimensional polymer liquid crystal. *Macromolecules* **29**, 8473–8478 (1996).
- Maruyama, T., Fuller, G. G., Grosso, M. & Maffettone, P. L. The dynamics of two dimensional polymer nematics. *J. Non-Newtonian Fluid Mech.* **76**, 233–247 (1998).
- Jeffery, G. B. Motion of ellipsoidal particles immersed in a viscous fluid. *Proc. R. Soc. Lond. A* **102**, 161–179 (1922).
- Nelson, D. R. & Halperin, B. I. Solid and fluid phases in smectic layers with tilted molecules. *Phys. Rev. B* **21**, 5312–5328 (1980).
- Selinger, J. V., Wang, Z.-G., Bruinsma, R. F. & Knobler, C. M. Chiral symmetry breaking in Langmuir monolayers and smectic films. *Phys. Rev. Lett.* **70**, 1139–1142 (1993).

Acknowledgements

This work was supported by the National Science Foundation, the donors of the Petroleum Research Fund, and the Camille & Henry Dreyfus Foundation.

Correspondence and requests for materials should be addressed to D.K.S. (e-mail: daniel.schwartz@colorado.edu).

Evolution of leaf-form in land plants linked to atmospheric CO₂ decline in the Late Palaeozoic era

D. J. Beerling*, C. P. Osborne* & W. G. Chaloner†

* Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

† Department of Geology, Royal Holloway, University of London, Egham, Surrey TW20 0EK, UK

The widespread appearance of megaphyll leaves, with their branched veins and planate form, did not occur until the close of the Devonian period at about 360 Myr ago. This happened about 40 Myr after simple leafless vascular plants first colonized the land in the Late Silurian/Early Devonian^{1,2}, but the reason for the slow emergence of this common feature of present-day plants is presently unresolved. Here we show, in a series of quantitative analyses using fossil leaf characters and biophysical principles, that the delay was causally linked with a 90% drop in atmospheric p_{CO_2} during the Late Palaeozoic era^{3,4}. In contrast to simulations for a typical Early Devonian land plant, possessing few stomata⁵ on leafless stems, those for a planate leaf with the same stomatal characteristics indicate that it would have suffered lethal overheating, because of greater interception of solar energy and low transpiration. When planate leaves first appeared in the Late Devonian and subsequently diversified in the Carboniferous period, they possessed substantially higher stomatal densities⁶. This observation is consistent with the effects of the p_{CO_2} on stomatal development⁷ and suggests that the evolution of planate leaves could only have occurred after an increase in stomatal density, allowing higher transpiration rates that were sufficient to maintain cool and viable leaf temperatures.

The origin of megaphyll leaves was an important event in land plant evolution with a major impact on a wide variety of terrestrial biogeochemical and ecological processes. Indeed, planate leaves are the basis of primary productivity for virtually all terrestrial life, especially tetrapods and insects. The fossil record unequivocally demonstrates that early vascular land plants in the Late Silurian/Early Devonian were either leafless, possessing short, cylindrical, aerial stems or bore only minute enation leaves (microphylls)². Over the next 40 Myr a progressive evolutionary sequence followed. The early axial-type plants showed a gradual shift, developing much-

branched determinate lateral systems ('proto-leaves') and then fully 'webbed' (that is, laminate) planate leaves with dichotomous venation by the Late Devonian¹. The long delay in this relatively simple evolutionary development is surprising given the ubiquity of the laminate leaf form in modern plant groups, its importance in photosynthetic carbon acquisition, and that the same interval of time witnessed the structurally far more complex evolution of vascular plant life cycles from homospory to the seed habit^{2,8}. This critical, but neglected, feature of land plant evolutionary biology has remained enigmatic since it was reported descriptively as the 'telome theory' over 70 years ago⁹. Megaphyll planate leaves became abundant in Late Devonian/Early Carboniferous fossilized terrestrial floras¹⁰, but first appeared in the Early Devonian plant *Eophyllophyton bellum*¹¹, indicating that the production of laminate leaves was possible, although not widespread, at this time.

Here we address this evolutionary problem by simulating organ gas exchange rates and energy budgets (see Methods) to quantitatively assess the probable costs and benefits of evolving a planate leaf. We first consider a simple erect axis and planate leaf operating in the Early Devonian atmosphere (Fig. 1), both with the low stomatal density characterizing plant fossils of this age (Table 1). Planate leaves at low latitudes intercept over 200% more solar energy than axes and, coupled with high stomatal resistance curtailing transpiration, this raises leaf temperatures well into the range for lethal damage observed in extant tropical taxa (50–55 °C)¹². Even at higher latitudes, where the solar angle is lower, temperatures reach 40 °C, a value close to the limit for photosynthetic CO₂ uptake¹². In contrast, an axis maintains low transpiration rates but avoids overheating by minimizing the surface area presented to intercept solar energy during the middle of the day. Given that axis rates of photosynthesis exceed those of surviving planate leaves, no selective advantage in terms of carbon gain is conferred by evolving this leaf form. Even if planate leaves had evolved with a high stomatal density, the resulting transpiration rates (9–13 mmol H₂O m⁻² s⁻¹) would only have been sufficient to cool leaf temperatures to 38 °C at low latitudes. These rates require xylem conductances of over ten times those measured¹³ for *Psilotum nudum*, a plant possessing a primitive stele with close anatomical similarities to early Rhyniophytes. Such transpiration rates are clearly therefore incompatible with the functioning of early land plants. However, the higher rates for an erect photosynthetic axis fall well within its hydraulic constraints.

The earliest preserved fossil megaphyll leaf cuticles date to the Carboniferous and have higher stomatal densities than early land plant axes by one to two orders of magnitude (Table 1). Geochemical models show a major draw-down in p_{CO_2} between the Early

Table 1 Leaf and environmental characteristics used in biophysical modelling

Leaf characteristics					
Type of leaf	Stomatal density* (mm ⁻²)	Stomatal pore length/width (µm)	Diameter of axis/leaf width (mm)	Diffusion path length (µm)†	$V_{\text{max}}/J_{\text{max}}$ (µmol m ⁻² s ⁻¹)‡
Low-density axis	10	40/20	4	500	11/48
Low-density planar leaf	10	40/20	60	100	11/48
High-density planar leaf	1,000	20/10	60	100	78/157
Midday palaeoenvironmental conditions§					
Environment	Atmospheric CO ₂ (Pa)¶	Atmospheric O ₂ (kPa)¶¶	Air temperature (°C)#		
			20° S	60° S	
Early Devonian	300	21	35	30	
Late Devonian/Carboniferous	30	30	35	30	

* The low values are representative of Late Silurian/Early Devonian plant fossils, and the high values are representative of Carboniferous plant fossils^{5,6,27}.

† Taken as 0.125 of axis diameter and 0.5 of leaf thickness, assuming a hypostomatous leaf.

‡ From ref. 22, where V_{max} and J_{max} are the maximum rates of carboxylation activity and photosynthetic electron transport respectively²³.

§ All simulations were made for conditions representative of still clear summer days with solar energy interception values calculated as in Methods, a relative humidity of 60% and a wind speed of 0.5 m s⁻¹, a value for conditions close to the ground surface, where small Early Devonian axes would have evolved planate leaves.

¶ From ref. 3.

¶¶ From ref. 14.

From land-surface summer temperature computed by a general circulation model for a pole-centred supercontinent in the Carboniferous³⁰. We have conservatively assumed no warming of summer temperatures in the high-CO₂ Early Devonian atmosphere, compared to the low-CO₂ Carboniferous atmosphere, and have not accounted for a diurnal temperature increase over mean monthly global climate model temperatures.

Devonian and Late Devonian/Early Carboniferous and a rise in the atmospheric p_{O_2} (refs 3, 14), both associated with increased organic carbon burial by the spread of Devonian forests, the formation of Carboniferous swamp lands, and plant-enhanced chemical weathering of rocks. Simulations for Late Devonian/Carboniferous leaves (Table 1) indicate transpiration rates sufficient to cool leaf temperatures well below the lethal range, even in the warm low latitudes (Fig. 1). The reduced stomatal resistance of these planate leaves benefited photosynthetic productivity under the unusually low atmospheric CO_2/O_2 ratio by increasing CO_2 diffusion into leaf mesophyll and thereby minimizing photorespiratory CO_2 evolution through the carbon oxidation pathway¹⁵.

We tested our biophysical simulations of ancient land plants by first computing their water-use efficiencies (carbon gain per unit of water lost) using modelled gas exchange characteristics. These were next compared with independent estimates obtained using the stable carbon isotope composition ($\delta^{13}\text{C}$) of fossilized Palaeozoic terrestrial organic matter⁴ and a well-validated model¹⁶ relating plant water-use efficiency to $\delta^{13}\text{C}$ (see Methods). Calculated in this way, both approaches consistently show similar values and a pattern of declining plant water-use efficiency between the Late Silurian period (410–415 Myr ago) and Carboniferous (300–292 Myr ago) (Fig. 2). Agreement with fossil evidence therefore provides qualitative and quantitative support for our modelled fluxes of CO_2 and H_2O between photosynthetic organs and the surrounding atmosphere.

Development of photosynthetic structures with high stomatal

densities, and correspondingly higher transpirational cooling capacities, was probably an essential requirement for the evolution of laminate leaves. Observations on a range of plant species show that atmospheric p_{CO_2} regulates leaf stomatal development, with the potential for genetic adaptation on a timescale of millions of years¹⁷. Therefore, the drawdown in atmospheric p_{CO_2} in the Late Palaeozoic era and the concurrent observed increase in stomatal density, is a likely ancient example of this effect of CO_2 on plant development. In this context, however, the increased stomatal density had important consequences for their subsequent morphological evolution. Moreover, a 40-Myr delay between the axial form of Late Silurian/Early Devonian land plants and the development of megaphyll planate leaves is consistent with the timescale required to remove CO_2 from the atmospheric reservoir by silicate rock weathering and organic carbon burial³.

The likelihood of high-temperature injury and collapse of photosynthetic productivity in Early Devonian plants will have been critically dependent upon the ultimate lobe width of planate photosynthetic structures, because this strongly influences boundary layer resistance and convective heat dissipation, that is, cooling capacity¹⁸. We calculate that widths of 50–100 mm at high latitudes and less than 20 mm in equatorial regions are maximal before temperatures increase and photosynthesis drops abruptly (Fig. 3). The earliest 'proto-leaves' should therefore have been rather narrow, especially for plants distributed in warm tropical regions, and this model prediction is supported by two pieces of evidence from the fossil record. Observations made on Early Devonian (Emsian) plant fossils at a palaeolatitude of $\sim 15^\circ\text{S}$ showed that some of the earliest megaphyllous, but non-laminate, leaves were finely divided (7–10 mm in width)¹⁹. The earliest yet reported laminate megaphyllous leaves with branched venation were very small and finely divided (up to 2–5 mm broad)¹¹, dated to the Early Devonian (Pragian);

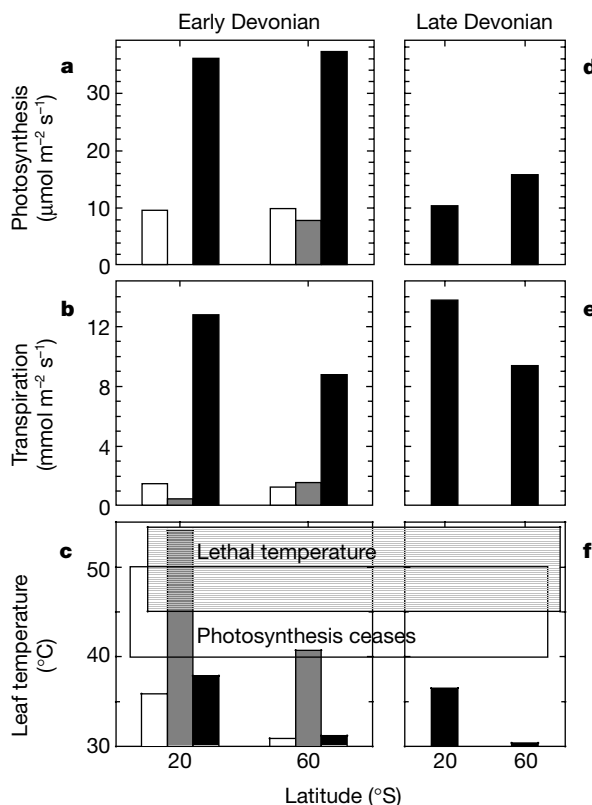


Figure 1 Simulated biophysical properties of leaves and axes in the Late Palaeozoic environment. Midday photosynthetic rates (**a**), transpiration rates (**b**) and temperatures (**c**) for an axis with a low (white) stomatal density and planate leaves with high (black) or low (grey) stomatal density operating in a high- p_{CO_2} Early Devonian atmosphere (Table 1) at 60°S and 20°S . The absence of the low stomatal density planate leaf in **a** is because its temperature is lethal (see **c**). **d–f**, Adjacent boxes show the same ecophysiological traits for a planate leaf with a high stomatal density in the lower p_{CO_2} atmosphere of the Late Devonian. Horizontal boxes in **c** and **f** indicate the temperature range at which CO_2 uptake by photosynthesis ceases and the lethal range for extant warm temperate and tropical species¹².

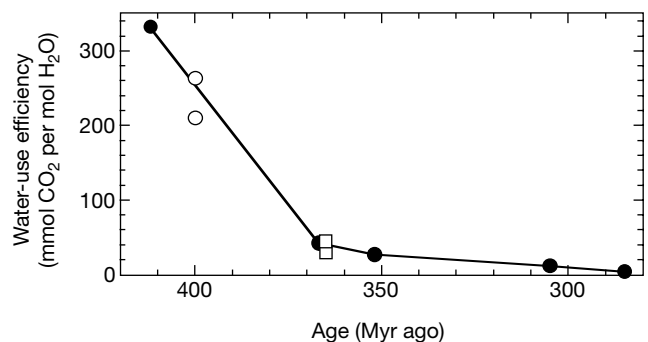


Figure 2 Stimulated and isotopically derived changes in leaf and axis water-use efficiency during the Late Palaeozoic era. Black circles, estimate from soil organic $\delta^{13}\text{C}$; white circles, simulated Early Devonian axis; white squares, simulated Early Carboniferous planate leaf.

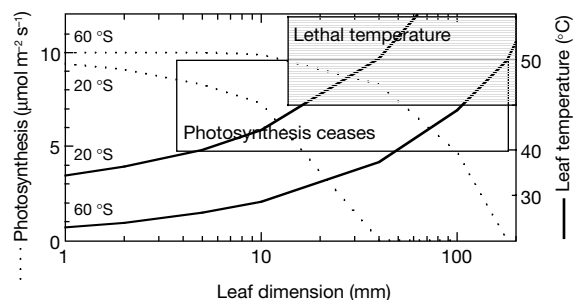


Figure 3 Simulated biophysical properties of Early Devonian planate leaves (Table 1) in relation to width. Dotted line, photosynthesis; solid line, leaf temperature, shown for two palaeolatitudes.

they belonged to *E. bellum* growing at a palaeolatitude of $\sim 25^\circ$ N. It follows that the energetic constraints precluding the evolution of wider laminate leaves would have been more relaxed in the cooler climates of the mid- to high latitudes, pointing to these regions as the likely geographical sources for this leaf form.

As the atmospheric p_{CO_2} fell during the Late Palaeozoic, we expect the laminate leaf to have become more widespread in the plant kingdom and progressively larger or less dissected. By the close of the Devonian, plants from several groups possessed planate leaves with an extensive dichotomizing venation. Some of these had megaphyll leaves up to 30–40 mm across, with a morphology similar to *Ginkgo*, and are assigned to several ill-defined genera including *Ginkgophyllum*, *Platyphyllum*, and *Psymphyllum*. The genus *Enigmophyton* Hoeg., probably of Middle Devonian age, had *Ginkgo*-like leaves up to 160 mm long and 120 mm broad. These examples constitute the earliest known occurrences of sizable laminate leaves in the fossil record¹ and clearly indicate that after 40 Myr, biotic and abiotic changes allowed a greater interception of solar energy without causing lethal temperature damage to the leaf, or significant high-temperature limitation to photosynthesis. An important consequence for terrestrial plants able to support planate leaves with high transpiration rates (Fig. 2) would have been the selective pressure driving improved water and nutrient transport, crucial for plants as their heights rapidly increased from the Middle Devonian onwards¹⁰. However, by the Middle to Late Devonian, fossilized plant assemblages reveal that vascular systems were probably more efficient²⁰, and water uptake from the soil improved by the evolution of deep rooting systems²¹. □

Methods

Simulations were made using a fully coupled photosynthesis/stomatal resistance/energy balance model, accounting for the feedbacks between changes in the environment and their influence on gas exchange²². Fossil stomatal geometry was used to calculate minimum stomatal resistance. We then allowed for the effects of CO_2 , leaf-to-air vapour pressure deficit, irradiance and temperature on stomatal pore width, and used this modified stomatal resistance to calculate the leaf energy budget. Photosynthetic rate was predicted with a steady state biochemical model of leaf CO_2 assimilation²³, with a feedback modifying the intercellular CO_2 concentration of the leaf which, in turn, influenced stomatal opening and thus changes in leaf energy balance.

V_{max} and J_{max} (Table 1) were derived by calculating the maximum rate of photosynthesis and corresponding values of intercellular CO_2 level from stomatal resistance using the fossil stomatal characteristics^{24,25}. In this way, the entire process of modelling leaf gas exchange was based on the fossil record and was independent of physiological initialization using data from plants growing in the present climate and CO_2 . These estimates are necessarily approximate, but have no effect on our conclusions regarding thermal damage to leaves in the Early Devonian. We note that the gene coding sequence, protein structure and kinetic properties of the key photosynthetic enzyme Rubisco are highly conserved among disparate taxonomic groups of C_3 plants²⁶, indicating little or no change in its properties throughout plant evolutionary history.

Within the coupled model, generalized responses of stomatal pore widths to the vapour pressure deficit, CO_2 and irradiance were used and, although there is some uncertainty attached to extrapolating these back to extinct plants, they are nevertheless representative of a wide range of modern taxa. All plants avoid desiccation under dry air conditions and in bright sunlight by closure of stomatal pores, as included in our model. In addition, anatomical studies of the guard cells of Devonian axes have shown that their form is remarkably close to that in modern plants²⁷, which suggests that they probably operated in a similar manner. Influences of the environment derived in the model from today's experiments simply modify the minimum stomatal resistance, calculated from the geometry and density of fossil stomata, rather than setting stomatal resistance itself.

Interception of solar energy by photosynthetic organs was computed from geometric considerations, with the zenith angle of planate leaves taken from reconstructions of permineralized Devonian plants ($\sim 40^\circ$ from the horizontal)¹ and integrated for all azimuth angles²⁸. Our model accounts for the greater distance travelled by CO_2 molecules from the atmosphere to the sites of carboxylation in an axis compared to a planate leaf (Table 1) by calculating total axis/leaf resistance to CO_2 diffusion (r_t , in units of s m^{-1}) using:

$$r_t = r_s + r_i + r_w + r_a$$

where r_s is the stomatal resistance to CO_2 diffusion of the leaf/axis, r_i and r_w are the diffusion resistances through the intercellular spaces, and across the cell wall to the site of photosynthesis (including plasmalemma, cytosol, mesophyll and chloroplast membrane) respectively. r_i is dependent upon the length of the diffusion pathway from the atmosphere to the site of photosynthesis (see Table 1). Because of the obvious difficulty of calculating r_w from fossil materials, we assumed a constant value for both axes and laminate leaves from measurements on modern plants of 240 s m^{-1} (ref. 18). The final term, r_a , is the boundary layer resistance which depends on wind speed and leaf width¹⁸ (see Table 1 for

leaf dimensions and environmental conditions). The calculated value of r_t consists of a combination of data obtained from measurements on fossil plant materials and assumed characteristics from extant plant leaves. However, r_t is dominated by r_a and r_s , and these are the two components best constrained by measurements on fossils.

The water-use efficiency of Palaeozoic plants was calculated with the isotopic composition of atmospheric CO_2 used during photosynthesis ($\delta^{13}\text{C}_a$), and measurements of the isotopic composition of terrestrial organic matter ($\delta^{13}\text{C}_b$)⁴, to calculate discrimination against ^{13}C (Δ) from $(\delta^{13}\text{C}_a - \delta^{13}\text{C}_b)/(1 + \delta^{13}\text{C}_b/1,000)$ (ref. 16), after correcting $\delta^{13}\text{C}_b$ to account for the difference between litter and leaf tissue values ($\sim 3\text{‰}$)²⁹. Plant water-use efficiency was calculated as $p_{\text{CO}_2} \times [1 - (-\Delta + a)/(-b + a)]/1.6$ where a is fractionation associated with diffusion (4.4‰), b is fractionation associated with Rubisco (27‰) and 1.6 is the ratio of gaseous diffusivities of CO_2 and water vapour in the air¹⁶. Modelled water-use efficiencies were calculated from the ratio of intercellular to atmospheric p_{CO_2} (c_i/c_a ratio) derived from the gas exchange data (Fig. 1) and used to calculate Δ as $a + (b - a) \times c_i/c_a$. This value was then used to calculate the water-use efficiency as for the carbon isotope data.

Received 29 August 2000; accepted 15 January 2001.

- Gensel, P. G. & Andrews, H. N. *Plant Life in the Devonian* (Praeger Scientific, New York, 1984).
- Kenrick, P. & Crane, P. R. The origin and early evolution of plants on land. *Nature* **389**, 33–39 (1997).
- Berner, R. A. GEOCARBII: a revised model of atmospheric CO_2 over Phanerozoic time. *Am. J. Sci.* **294**, 56–91 (1994).
- Mora, C. I., Driese, S. G. & Colarusso, L. A. Middle to late Paleozoic atmospheric CO_2 levels from soil carbonate and organic matter. *Science* **271**, 1105–1107 (1996).
- Edwards, D. Climate signals in Palaeozoic land plants. *Phil. Trans. R. Soc.* **353**, 141–157 (1998).
- McElwain, J. C. & Chaloner, W. G. Stomatal density and index of fossil plants track atmospheric carbon dioxide in the Palaeozoic. *Ann. Bot.* **76**, 389–395 (1995).
- Woodward, F. I. Stomatal numbers are sensitive to CO_2 increases from pre-industrial levels. *Nature* **327**, 617–618 (1987).
- Chaloner, W. G. & Hemsley, A. R. in *Pollen and Spores* (eds Blackmore, S. & Barnes, S. H.) 151–167 (Clarendon, Oxford, 1991).
- Zimmermann, W. *Die Phylogenie der Pflanzen* (Jena, Fischer, 1979).
- Chaloner, W. G. & Sheerin, A. in *The Devonian System* (eds House, M. R., Scrutton, C. T. & Bassett, M. G.) 145–161 (Palaeontological Society Special Paper in Palaeontology, no. 23, 1979).
- Hao, S. G. & Beck, C. B. Further observations on *Eophyllophyton bellum* from the lower Devonian (Siegenian) of Yunnan, China. *Palaeontographica B* **230**, 27–47 (1993).
- Larcher, W. in *Ecophysiology of Photosynthesis* (eds Schultze, E. D. & Caldwell, M. M.) 261–277 (Springer, Berlin, 1994).
- Schulte, P. J., Gibson, A. C. & Nobel, P. S. Xylem anatomy and hydraulic conductance of *Psilotum nudum*. *Am. J. Bot.* **74**, 1438–1445 (1987).
- Berner, R. A. et al. Isotope fractionation and atmospheric oxygen: implications for Phanerozoic O_2 evolution. *Science* **287**, 1630–1633 (2000).
- Beerling, D. J. et al. The influence of Carboniferous palaeo-atmospheres on plant function: an experimental and modelling assessment. *Phil. Trans. R. Soc.* **353**, 131–140 (1998).
- Farquhar, G. D., Ehleringer, J. R. & Hubrick, K. T. Carbon isotope discrimination and photosynthesis. *Ann. Rev. Plant Physiol. Plant Mol. Biol.* **40**, 503–537 (1989).
- Beerling, D. J. & Chaloner, W. G. Evolutionary responses of stomatal density to global CO_2 change. *Biol. J. Linn. Soc.* **48**, 343–353 (1993).
- Nobel, P. S. *Physicochemical and Environmental Physiology* (Academic, San Diego, 1991).
- Gensel, P. G. A new lower Devonian plant and the early evolution of leaves. *Nature* **309**, 785–787 (1984).
- Knoll, A. H. et al. Character diversification and patterns of evolution in early vascular plants for use in dynamic vegetation models. *Paleobiology* **10**, 34–47 (1984).
- Algeo, T. J. & Scheckler, S. E. Terrestrial-marine teleconnections in the Devonian: links between the evolution of land plants, weathering processes and marine anoxic events. *Phil. Trans. R. Soc.* **353**, 113–130 (1998).
- Beerling, D. J. & Woodward, F. I. Changes in land plant function over the Phanerozoic: reconstructions based on the fossil record. *Bot. J. Linn. Soc.* **124**, 137–153 (1997).
- Farquhar, G. D., Von Caemmerer, S. & Berry, J. A. A biochemical model of photosynthetic CO_2 assimilation in leaves of C_3 species. *Planta* **149**, 78–90 (1980).
- Von Caemmerer, S. & Farquhar, G. D. Some relationships between the biochemistry of photosynthesis and the gas exchange of leaves. *Planta* **153**, 376–387 (1981).
- Beerling, D. J. & Quick, W. P. A new technique for estimating rates of carboxylation and electron transport in leaves of C_3 plants for use in dynamic vegetation models. *Glob. Change Biol.* **1**, 289–294 (1995).
- Bowes, G. Facing the inevitable—plants and increasing atmospheric CO_2 . *Ann. Rev. Plant Physiol. Plant Mol. Biol.* **44**, 309–332 (1993).
- Edwards, D., Kerp, H. & Hass, H. Stomata in early land plants: an anatomical and ecophysiological approach. *J. Exp. Bot.* **49**, 309–332 (1993).
- Niklas, K. J. The role of phyllotactic pattern as a “developmental constraint” on the interception of light by leaf surfaces. *Evolution* **42**, 1–16 (1988).
- Martinelli, L. A. et al. Stable carbon isotope ratio of tree leaves, boles and fine litter in a tropical forest in Rondonia, Brazil. *Oecologia* **114**, 170–179 (1998).
- Valdes, P. J. & Crowley, T. J. A climate model intercomparison for the Carboniferous. *Palaeoclimates* **2**, 219–238 (1998).

Acknowledgements

We thank D. Edwards, J. A. Raven, F. I. Woodward and G. R. Upchurch for helpful comments and discussion on the manuscript. D.J.B. gratefully acknowledges funding through a Royal Society University Research Fellowship and the Natural Environment Research Council, UK.

Correspondence and requests for materials should be addressed to D.J.B. (e-mail: d.j.beerling@sheffield.ac.uk).

Increases in greenhouse forcing inferred from the outgoing longwave radiation spectra of the Earth in 1970 and 1997

John E. Harries, Helen E. Brindley, Pretty J. Sagoo & Richard J. Bantges

Space and Atmospheric Physics Group, Blackett Laboratory, Imperial College, London SW7 2BW, UK

The evolution of the Earth's climate has been extensively studied^{1,2}, and a strong link between increases in surface temperatures and greenhouse gases has been established^{3,4}. But this relationship is complicated by several feedback processes—most importantly the hydrological cycle—that are not well understood^{5–7}. Changes in the Earth's greenhouse effect can be detected from variations in the spectrum of outgoing longwave radiation^{8–10}, which is a measure of how the Earth cools to space and carries the imprint of the gases that are responsible for the greenhouse effect^{11–13}. Here we analyse the difference between the spectra of the outgoing longwave radiation of the Earth as measured by orbiting spacecraft in 1970 and 1997. We find differences in the spectra that point to long-term changes in atmospheric CH₄, CO₂ and O₃ as well as CFC-11 and CFC-12. Our results provide direct experimental evidence for a significant increase in the Earth's greenhouse effect that is consistent with concerns over radiative forcing of climate.

Starting in October 1996, the Interferometric Monitor of Greenhouse Gases (IMG) instrument¹⁴, on board the Japanese ADEOS satellite, produced about nine months of global observations of the spectrum of outgoing longwave radiation. Some 27 years earlier, NASA had flown a similar instrument (IRIS—Infrared Interferometric Spectrometer¹⁵) on the Nimbus 4 spacecraft, between April 1970 and January 1971. Details of the performance of the two instruments are given in Table 1. Although these two experiments were on board separate spacecraft almost 30 years apart, the existence of the two data sets allows the examination of the outgoing longwave radiation spectra to detect whether any significant change had occurred between the two measurements. Comparing averaged spectra for the same region, allowing for the different spatial and spectral resolutions of the two instruments, reveals that they are extremely similar and highly reproducible. Nevertheless, on closer inspection, informative differences are observed.

To illustrate this, Fig. 1a shows an averaged IRIS cloud-cleared brightness temperature spectrum superposed on an averaged IMG cloud-cleared spectrum, for the same three-month period (April–June) and a central Pacific area (latitude 10° N–10° S, longitude 130° W–180° W) for the range 710–1,400 cm⁻¹. The spectral resolution of the IMG data has been degraded to match that of IRIS, instrument field-of-view effects have been taken into account, and land/island areas have been masked out. For cloud clearing, a two-step approach was used¹⁶. Figure 1b shows three curves: upper, the difference IMG–IRIS taken directly from Fig. 1a; middle, a theoretical difference spectrum, corresponding to conditions for the central Pacific; and lower, the observed difference spectrum for a 'near-global' case (60° N–60° S) for the same period, for comparison. The curves in Fig. 1b are displaced by 5 K from one another. Figure 1c shows the component of the simulated spectrum that includes only the effect of trace-gas changes between 1970 and 1997 (omitting temperature and humidity changes), to aid interpretation.

The agreement between the upper and middle curves of Fig. 1b is very good. This result, given that the IMG and IRIS data were

recorded using two separate spectrometers, 27 years apart, and that the simulation is completely independent of IRIS/IMG observations, gives confidence in the quality of the data. The internal consistency of the data is further illustrated by the inclusion of the bottom, quasi-global result. Despite the fact that the averages in the top and bottom curves contain very different numbers of spectra, the consistency between these results is notable: the major features of the observed difference spectra appear consistently in all the difference spectra we have studied.

Our interpretation of Fig. 1 is as follows. We consider first the sharp spectral features. A negative-going brightness temperature difference is observed on the edge of the CO₂ ν_2 band, between 710 and 740 cm⁻¹, in accord with the known increase in atmospheric CO₂ concentrations between 1970 and 1997¹. The O₃ band centred at about 1,060 cm⁻¹ also shows a negative-going difference from the background window signal, which can be attributed to the known changes in ozone¹⁷ and in temperature¹⁸. A strong, negative Q-branch is observed at 1,304 cm⁻¹ in the CH₄ band, due mainly to increases in tropospheric CH₄ concentrations in the period between the observations, which causes emission from higher, colder layers of the troposphere. Negative-going lines due to ν_2 -band H₂O absorption are seen between 1,200 and 1,400 cm⁻¹. There is also evidence of weak features due to CO₂, CFC-11 and CFC-12 in the 700–1,000 cm⁻¹ range.

We now consider the background difference spectrum in the two window regions on either side of the O₃ band. The window difference spectrum could be influenced by changes in surface temperature (assuming constant oceanic emissivity), humidity of the lower troposphere, aerosol content, or cloud amount and type. Consistently, the difference in the 1,100–1,200 cm⁻¹ window is close to zero (within ± 1 K), while the difference in the 800–1,000 cm⁻¹ region is positive, and lies between about 1 and 2 K. It is important not to over-interpret the observations to an accuracy that is not justified by the errors (see below), nor to lose sight of our principal result, which is the observation of the sharp spectral features discussed in the preceding paragraph. Nevertheless, we believe that this reproducible difference between the windows is consistent with small residual amounts of ice cloud in both averaged spectra, possibly exacerbated by the different fields of view of the two instruments: we discuss this further below.

The simulations shown in Fig. 1b and c were calculated as follows. Profiles of atmospheric temperature and water vapour were extracted covering the same region and three-month time periods from the NCEP (National Centers for Environmental Prediction, Washington) reanalysis project¹⁹. Stratospheric ozone changes were estimated using measured trends²⁰ extrapolated back to 1970, whereas tropospheric ozone changes were calculated using a three-dimensional chemical transport model, forced by realistic emission scenarios²¹. Remaining gaseous concentrations were taken from the relevant IPCC values for CO₂, CH₄, N₂O, CFC-11 and CFC-12, in 1970 and 1997. The MODTRAN3 code²² was used to calculate the expected radiance spectra in 1970 and 1997.

All the principal features due to changes in CO₂, CH₄, O₃, temperature and humidity are well modelled, as are the small

Table 1 IMG and IRIS instrument characteristics

Characteristic	IMG	IRIS
Spectral range (cm ⁻¹)	600–3,000	400–1,600
Spatial field of view	8 km × 8 km	100 km × 100 km
Spectral resolution (cm ⁻¹)	0.10–0.25	2.8
Total radiometric uncertainty (mW m ⁻² sr ⁻¹ per cm ⁻¹)	±0.23	±(0.5†–1.0‡)
Equivalent brightness temperature uncertainty* (K)	±0.15	±(0.3†–0.6‡)

* Defined at 900 cm⁻¹ for a brightness temperature of 290 K.

† In centre of band.

‡ Towards edge of band.

changes due to the chlorofluorocarbons (for example, at 850 and 920 cm^{-1}) and weak CO_2 bands (for example, at 795 cm^{-1}). We note that the main features of the observed difference spectrum can only be reproduced by including the long-term changes in CH_4 , CO_2 , O_3 and the chlorofluorocarbons: inter-annual and short-term variability is not sufficient.

We must consider, however, whether any aspect of the observed difference spectra could be caused by instrumental, or other, artefacts. We have considered three: the accuracy and precision of the spectra; possible causes of differential window signals; and inter-annual variability.

First, we have examined the central issue of the accuracy and precision of the two data sets, and whether our analysis is justified on this count. We believe it is, for the following reasons. (1) We have reported above the agreement between the observed difference spectra and simulated spectra, which are calculated quite independently from basic knowledge of the atmospheric state. (2) We have derived difference spectra over a wide variety of regions and times (including the east and west Pacific, Atlantic, and Indian oceans) and see consistency in the principal absorption features. (3) Random error is reduced by averaging: in the extreme case in which all the error in Table 1 is assumed to be random, we obtain a value of ± 0.058 K for the central Pacific difference spectra. (4) Absolute accuracy is potentially a more serious issue: if all the error quoted in Table 1 were systematic, this would indicate a maximum absolute

peak–peak error for the combined data of about 0.45 K at the centre of the IRIS/IMG pass bands, increasing to about 0.75 K at the edges. Such errors of absolute calibration vary slowly with wavenumber, and could produce small ‘ghosts’ at the positions of the observed spectral features, though these are very unlikely to be significant. Slowly varying systematic errors²³ may also contribute to the differential window signal, but only at this same level of <0.45 K. (5) Normalization of each averaged spectrum to the intensity at one selected wavenumber indicates no significant multiplicative error between IMG and IRIS above the 0.5 K level. (6) Examination of the data shows that neither spatial nor temporal sampling is seriously biased. We conclude that the main features of the observed spectra cannot be accounted for by instrument errors, but that the absolute calibration of the difference spectra might be in error by up to about 0.5 to 0.75 K peak–peak.

Second, we have considered the observed differences in the two window background regions, and the influence of the different fields of view of the two instruments. (We point out that this influences only the background levels: we have averaged differing numbers of IRIS and IMG spectra to show that the main features in the difference spectrum due to greenhouse gases are not dependent on the fields of view of the two instruments). However, broad-band difference signals could occur if aerosol or cloud ‘contamination’ remains in the notionally clear fields of view. Using available aerosol data²⁴, we have shown for typical conditions that aerosol absorption is unlikely to be a significant source of error (<0.1 K). However, recent work²⁵ has shown that ice cloud, particularly if composed of small crystals, does exhibit stronger absorption in the 800–1,000 cm^{-1} than the 1,100–1,200 cm^{-1} window. It is quite possible that small residual amounts of ice cloud absorption remain in both sets of data. Owing to the larger field of view, the IRIS spectra have a much higher probability of being contaminated than their IMG counterparts. The observed 1 K or so enhancement of the 800–1,000 cm^{-1} difference signal would be consistent with this, and could also arise from a change in the mean cirrus microphysical properties. We cannot separate these two effects, but we do conclude that the observed window difference spectra strongly indicate an effect involving residual small ice crystal effects, incompletely cleared from the data. R.J.B. has performed further calculations, following on earlier work²⁶, which confirm that window difference spectra of the magnitude observed can easily arise from small changes in amount, size or shape of small ice crystals: these studies also indicate that the difference spectrum should be larger below about 920 cm^{-1} , which is consistent with the observed data, especially the global case (Fig. 1b). Further work on these and other cloud effects in the data will be performed separately: for the present, we believe we have demonstrated a sufficient understanding of the observations to give confidence to the principal findings of this work regarding changes in radiative forcing due to CH_4 , CO_2 , O_3 and the chlorofluorocarbons.

Third, we must also take into account inter-annual variability as a possible cause of the observed difference spectra. In the window region, the brightness temperature difference is strongly modulated by short-term fluctuations, such as inter-annual variability (specific concern involves the 1997 warm El Niño/Southern Oscillation, ENSO, event). Our studies show that, while this could account for an uncertainty of about 1 K in the position of the zero line in the spatially and temporally averaged difference spectra used, it could not account for the sharp spectral features observed, nor the differential window signal just discussed.

The results presented here provide (to our knowledge) the first experimental observation of changes in the Earth’s outgoing long-wave radiation spectrum, and therefore the greenhouse effect: previous studies have been largely limited to theoretical simulations because of the paucity of data. We intend to examine the temporal and spatial variation of the difference spectra, and will include cloudy and all sky data. Future measurements by an IRIS-type

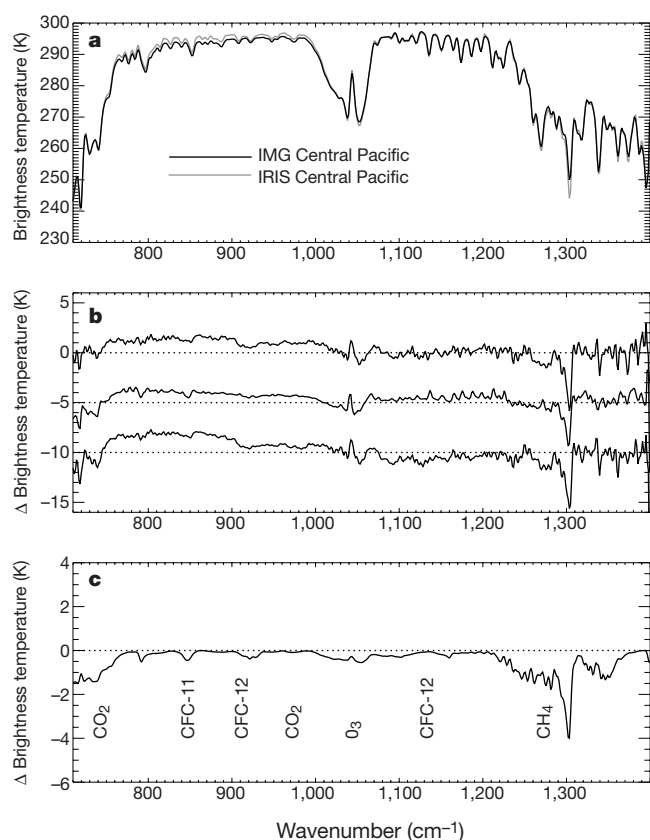


Figure 1 Examples of IRIS and IMG observed and simulated spectra for a three-month average (April–June) over selected regions. **a**, Observed IRIS and IMG clear sky brightness temperature spectra for the central Pacific (10°N–10°S, 130°W–180°W). **b**, Top, observed difference spectrum taken from **a**; middle, simulated central Pacific difference spectrum, displaced by -5 K; bottom, observed difference spectrum for ‘near-global’ case (60°N–60°S), displaced by -10 K. **c**, Component of simulated spectrum due to trace-gas changes only. ‘Brightness temperature’ on the ordinate indicates equivalent blackbody brightness temperature.

instrument that combines high accuracy and a narrow field of view are a priority for climate research. □

Methods

Reduction of IMG to IRIS spectral resolution and solid-angle correction

Before any further analysis, the IMG data were smoothed to match the IRIS resolution. In addition, the wavenumber scale in the two instruments is slightly different, owing to differing solid angles within the instruments²⁷. An empirically derived multiplicative factor of 0.9995 was applied to the IRIS wavenumber scale. Our work indicates that this process needs to be performed extremely accurately in order to avoid artefacts appearing in the difference spectrum.

Cloud clearing and spatial resolution

Cloud clearing was achieved in two steps¹⁶. First, for each spectrum, the brightness temperature at $1,126.6\text{ cm}^{-1}$ (surface to top of atmosphere transmittance ≈ 0.94) was compared to the underlying sea surface temperature taken from NCEP for the same location and time. If the difference exceeded a set threshold, the spectrum was rejected as being cloud-contaminated. The threshold for both IMG and IRIS was set at 6 K. This value was found by testing the variation of the standard deviation of the spectra with the threshold value.

The remaining spectra were subjected to a second filtering, to try to remove any residual spectral effects of cirrus clouds. The brightness temperature difference between two wavenumbers (909.8 and $1,250.4\text{ cm}^{-1}$) was compared with a threshold value¹⁶. The threshold value was chosen as before, by noting the temperature difference at which the standard deviation of the set of IRIS, or IMG, spectra fell to a near-constant value. A value of 8 K was used for IRIS spectra and 7 K for IMG. More severe cloud-clearing thresholds led to either all IRIS spectra being rejected as contaminated, or insufficient numbers of IMG spectra being retained to give the requisite spatial and temporal coverage.

Numbers of retained spectra in Fig. 1

The numbers of spectra retained and used in averaging to produce Fig. 1, after cloud clearance, were as follows. Global: IMG, 4,061; IRIS, 529. Central Pacific: IMG, 213; IRIS, 28.

Random and absolute errors

The total uncertainty of IRIS and IMG are quoted in Table 1. To examine the random and absolute components of this, we have taken the quoted errors to represent either the maximum random error (for example, caused by noise in the detector or electronics system) or the maximum systematic error (for example, multiplicative or additive error, such as absolute errors in the transmission or reflectivity of optical components such as filters and mirrors, or the temperature or emissivity of the blackbody targets used to calibrate the instruments, or an offset due to vignetting in the field of view). Such errors would typically be slowly varying with wavenumber.

For the random error, we have divided the errors quoted in Table 1 by the square root of the number of spectra left in each cloud-cleared average, and combined IMG and IRIS errors in a root-mean-square sense. The error in the difference spectrum of Fig. 1 amounts to $\pm 0.058\text{ K}$. For systematic error, from Table 1 the maximum peak–peak error that could arise, and which would not be reduced by averaging, would be 0.45 K in the band centre and 0.75 K towards the edge of the band.

We note that multiplicative errors of absolute calibration as described above could produce sharp features in a difference spectrum, but that these would not exceed the quoted peak–peak error. Thus we conclude that the maximum systematic error, slowly varying across the spectrum, is probably $\leq 0.5\text{ K}$.

Spectral range used in comparison

The upper limit of $1,400\text{ cm}^{-1}$ used in this analysis was based on the useful signal to noise ratio of IRIS; the lower limit of 710 cm^{-1} was based on the recommendations of the IMG Science Working Group.

Received 17 May 2000; accepted 15 January 2001.

- Houghton, J., Jenkins, G. & Ephraums, J. (eds) *Climate Change: The IPCC Scientific Assessment* (Cambridge Univ. Press, Cambridge, 1990).
- Houghton, J. et al. (eds) *Climate Change 1995: The Science of Climate Change* (Cambridge Univ. Press, Cambridge, 1995).
- Mitchell, J., Johns, T., Gregory, J. & Tett, S. Climate response to increasing levels of greenhouse gases and sulphate aerosols. *Nature* **376**, 501–505 (1995).
- Tett, S., Stott, P., Allen, M., Ingram, W. & Mitchell, J. Causes of twentieth century temperature change near the Earth's surface. *Nature* **399**, 569–575 (1999).
- Cess, R. et al. Interpretation of cloud-climate feedback predicted by 14 atmospheric general circulation models. *Science* **245**, 513–516 (1989).
- Rind, D. et al. Positive water vapour feedback in climate models confirmed by satellite observations. *Nature* **349**, 500–503 (1991).
- Lindzen, R. Some coolness regarding global warming. *Bull. Am. Meteorol. Soc.* **71**, 288–299 (1990).
- Goody, R., Haskins, R., Abdou, W. & Chen, L. Detection of climate forcing using emission spectra. *Earth Observ. Remote Sensing* **13**, 713–722 (1995).
- Harries, J., Brindley, H. & Geer, A. Climate variability and trends from operational satellite spectral data. *Geophys. Res. Lett.* **25**, 3975–3978 (1998).

- Goody, R., Anderson, J. & North, G. Testing climate models: an approach. *Bull. Am. Meteorol. Soc.* **79**, 2541–2549 (1998).
- Slingo, A. & Webb, M. The spectral signature of global warming. *Q. J. R. Meteorol. Soc.* **123**, 293–307 (1996).
- Clough, S., Iacono, M. & Moncet, J.-L. Line-by-line calculations of atmospheric fluxes and cooling rates: Application to water vapor. *J. Geophys. Res.* **97**, 15761–15785 (1992).
- Clough, S. & Iacono, M. Line-by-line calculation of atmospheric fluxes and cooling rates: 2. Application to carbon dioxide, methane, nitrous oxide and the halocarbons. *J. Geophys. Res.* **100**, 16519–16535 (1995).
- IMG mission operation and verification committee. *Interferometric Monitor for Greenhouse Gases* (ed. Kobayashi, H.) (IMG Project Technical Report, Central Research Institute of Electric Power Industry (CRIEPI), Komae Research Laboratory, Komae-shi, Tokyo, 1999).
- Hanel, R., Schlachman, B., Rogers, D. & Vanous, D. Nimbus 4 Michelson interferometer. *Appl. Opt.* **10**, 1376–1382 (1971).
- Strabala, K., Ackerman, S. & Menzel, W. Cloud properties inferred from 8–12 μm data. *J. Appl. Meteorol.* **33**, 212–229 (1994).
- Scientific Assessment of Ozone Depletion: 1998* Ch. 4 (Report No. 44, Global Ozone Research and Monitoring Project, World Meteorological Organization, Geneva, 1999).
- Oort, A. & Liu, H. Upper air temperature trends over the globe. 1958–1989. *J. Clim.* **6**, 292–307 (1993).
- Kalnay, E. et al. The NCEP/NCAR 40-year re-analysis project. *Bull. Am. Meteorol. Soc.* **77**, 437–471 (1996).
- Randel, W. & Wu, F. A stratospheric ozone trends data set for global modeling studies. *Geophys. Res. Lett.* **26**, 3089–3092 (1999).
- Olivier, J. et al. *Description of EDGAR Version 2.0: A Set of Global Emission Inventories of Greenhouse Gases and Ozone-depleting Substances for all Anthropogenic and Most Natural Sources on a Per Country Basis and on a 1 Deg x 1 Deg Grid* (RIVM Report No. 771060 002 and TNO-MEP Report No. R96/119, National Institute of Public Health and the Environment, Bilthoven, The Netherlands, 1996).
- Berk, A., Bernstein, L. & Robertson, D. *MODTRAN: a Moderate Resolution Model for LOWTRAN 7* (AFGL-TR-88-0177, US Air Force Geophysics Laboratory, Hanscom Air Force Base, Massachusetts, 1989).
- Hanel, R. et al. The Nimbus 4 infrared spectroscopy experiment 1. Calibrated thermal emission spectra. *J. Geophys. Res.* **77**, 2629–2641 (1972).
- Shettle, E. P. in *Atmospheric Propagation in the UV, Visible, IR and MM-wave Region and Related System Aspects* 15–1–15–12 (AGARD-CP-454, Air Force Geophysics Lab., Bedford, Massachusetts, 1990).
- Ackerman, S., Smith, W., Spinhirne, J. & Revercomb, H. The 27–28 October 1986 FIRE IFO cirrus case study: spectral properties of cirrus cloud in the 8–12 μm window. *Mon. Weath. Rev.* **118**, 2377–2388 (1990).
- Bantges, R., Russell, J. & Haigh, J. Cirrus cloud top-of-atmosphere radiance spectra in the thermal infrared. *J. Quant. Spectrosc. Radiat. Transfer* **63**, 487–498 (1999).
- Chamberlain, J. *The Principles of Interferometric Spectroscopy* (Wiley, Chichester, 1979).

Acknowledgements

We thank L. Chen and R. Goody for providing access to IRIS data, and S. Taguchi and H. Kobayashi for providing access to IMG data.

Correspondence and requests for materials should be addressed to J.E.H. (e-mail: j.harries@ic.ac.uk).

Fossil evidence of water lilies (Nymphaeales) in the Early Cretaceous

Else Marie Friis*, Kaj Raunsgaard Pedersen† & Peter R. Crane‡

* Department of Palaeobotany, Swedish Museum of Natural History, Box 50007, SE-104 05 Stockholm, Sweden

† Department of Geology, University of Aarhus, Universitetsparken, DK-8000 Århus C, Denmark

‡ Royal Botanical Gardens, Kew, Richmond, Surrey TW9 3AB, UK

Phylogenetic analyses have identified the water lilies (Nymphaeales: Cabombaceae and Nymphaeaceae), together with four other small groups of flowering plants (the 'ANITA clades': Amborellaceae, Illiciales, Trimeniaceae, Austrobaileyaceae), as the first diverging lineages from the main branch of the angiosperm phylogenetic tree^{1–4}, but evidence of these groups in the earliest phases of the angiosperm fossil record has remained elusive. Here we report the earliest unequivocal evidence, based

on fossil floral structures and associated pollen, of fossil plants related to members of the ANITA clades. This extends the history of the water lilies (Nymphaeales) back to the Early Cretaceous (125–115 million years) and into the oldest fossil assemblages that contain unequivocal angiosperm stamens and carpels. This discovery adds to the growing congruence between results from molecular-based analyses of relationships among angiosperms and the palaeobotanical record. It is also consistent with previous observations that the flowers of early angiosperms were generally very small⁵ compared with those of their living relatives.

Syntheses of the rapidly emerging fossil history of principal angiosperm clades produce a pattern of first appearances that is broadly congruent with the relative divergence times predicted from phylogenetic analyses of molecular data⁶. Fossils from the earliest phases of angiosperm diversification provide clear evidence of the family Chloranthaceae; however, unequivocal evidence of other clades of predicted greater or equal antiquity (for example, the ANITA clades and monocots) has so far been confined to the Late Cretaceous^{7,8}.

The fossil material described here is from the Vale de Agua locality, a large complex of clay pits situated in the Beira Littoral region of the western Portuguese Basin about 5 km southwest of Batalha. The fossil-bearing strata are mostly clays belonging to the 'Complexos gresosos de Nazaré e de Cós-Juncal'. The mesofossil assemblage from Vale de Agua comprises a rich flora of angiosperm reproductive organs, as well as many twigs and other remains of non-angiospermous seed plants and pteridophytes^{9,10}. About 100 different angiosperm taxa have been recognized in the flora, comprising flowers, fruits and seeds. About 30 different kinds of angiosperm pollen have also been recognized *in situ* within stamens and on the surfaces of fruits.

Strata at the Vale de Agua locality are of Early Cretaceous age and deposited in a terrestrial environment¹¹. The *in situ* palynological assemblage indicates that the strata are not younger than the Aptian, and possibly not older than the Barremian. Ninety per cent of the angiosperm pollen grains of the Vale de Agua flora are mono-aperturate, whereas only 10 per cent are tricolpate. The palynoflora thus correlates with Barremian–Aptian palynological assemblages from other parts of the world containing angiosperm pollen^{12–14}. Tricolpate pollen first appears in the fossil record in the Barremian¹⁵, but constitutes only a minor component of Barremian and early Aptian palynofloras, which are all strongly dominated by mono-aperturate pollen¹². Other triaperturate pollen types, such as tricolporoidate, tricolporate and triplicate forms that are common and sometimes dominant among the angiosperm component of Albian and younger pollen floras^{12,16}, are not present in the Vale de Agua assemblage.

Several key taxa among the angiosperm pollen assemblage that are typical for pre-Albian strata are also present in the Vale de Agua flora (such as forms comparable to Hauterivian *Retisulc-muriverm*, Barremian-*teebac*¹⁷, and early acolumellate–reticulate species¹³). On the basis of geographical proximity, similar geology and many shared elements of the mesofloras and pollen flora, the Vale de Agua assemblage is comparable in systematic composition and age to four other floras from terrestrial sediments in the western Portuguese Basin (Torres Vedras, Buarcos, Catefica, Famalicão)^{9,10}. According to a detailed sedimentological and palaeontological study¹⁸ of the Estremadura Region, the strata that include the Torres Vedras flora are not younger than the Aptian and are possibly Barremian or even older. Thus the combined palynological and geological data for the Vale de Agua assemblage strongly indicate a Barremian–Aptian age. Additional details of the locality, geological occurrence and composition of the Vale de Agua flora are provided elsewhere^{9,10}.

The single specimen (Fig. 1) is coalified, structurally preserved and only slightly compressed laterally. It was studied with light and scanning electron microscopy (SEM). The flower is well preserved,

probably fossilized at the time of anthesis, and minute—about 3 mm long and 2 mm in diameter, and even allowing for considerable shrinkage during fossilization the flower would not have exceeded 1 cm in diameter in life with all floral organs preserved. The flower is radially symmetrical, perigynous and apparently bisexual (Fig. 1a, b, d). All floral organs surrounding the carpels have broken off leaving irregular remnants of the organ bases. Twenty-four distinct rhombic remains (Fig. 1a, b), which we interpret as stamen bases (Fig. 1d), occur around the gynoecium and appear to form two whorls. The 12 inner stamen bases are opposite the carpels, and those of the outer whorl alternate with the carpels. The shape of the bases indicates that the stamen filaments were linear rather than laminar. Below the stamen bases there are six narrowly elliptical transverse organ bases, which we interpret as the bases of perianth parts (Fig. 1a, d). The arrangement of these organs is not clear, but may be spiral.

Pollen grains occur on the surface of the flower between the stamen scars. Grains are monocolpate and coarsely reticulate, about 20 µm in diameter (Fig. 1c), with muri of the reticulum supported by short columellae. Because the pollen is of a single kind and is restricted to the androecial zone, these grains are likely to have been produced by the stamens.

The gynoecium is syncarpous formed from 12 carpels arranged in a whorl around a central protrusion of the floral apex (Fig. 1a, b). Carpels are united below, and are free from each other only at the apex. The tips extend upwards and outwards to form a shallow depression with a short protrusion in its centre (Fig. 1b). Ventral sutures, which extend from the tip of the carpels to the central protrusion, are apparently open. Opening of the carpels may have occurred during fossilization, but the sutures may also have been open in life and closed only by secretion, as in some extant ANITA clades. The stigmatic area is indistinct and appears to consist of discrete, radiating, narrow, stigmatic zones on either side of the ventral suture of each carpel. Preservation is not sufficiently good to show whether the stigmatic zones extend to the tips of the carpels. A fracture (made after SEM) shows that each carpel contains several ovules, but their form, precise number and arrangement cannot be established with certainty.

Several features of the fossil flower clearly indicate a close relationship with extant Nymphaeales. The arrangement of carpels around a central protrusion of the floral apex is a unique feature that is so far known only for some members of the Nymphaeaceae and for the Illiciaceae¹⁹. However, flowers of Illiciaceae are hypogynous in contrast to the fossil flower, which is perigynous. Illiciaceae flowers also differ in having an apocarpous gynoecium and a single ovule per carpel. The fossil is more similar to flowers of Nymphaeales, which are predominantly syncarpous, include perigynous or epigynous forms in all genera except *Nuphar*, and always have more ovules per carpel.

The presence of Nymphaeales in the Vale de Agua fossil assemblage is also supported by the occurrence of many small exotestal seeds characterized by a testa composed of distinct palisade cells with strongly wavy anticlinal walls^{9,10}. Such seeds are characteristic of extant Nymphaeales and Illiciales. Leaf imprints of possible water lilies have also been described from the Early Cretaceous (Albian or older) of Portugal (*Braseniopsis*²⁰), and from the Early and mid-Cretaceous (late Aptian–Cenomanian) of Jordan (D. W. Taylor, G. J. Brenner, S. H. S. Basha and A. H. Al-Hammad, personal communication). However, none of these other Early to mid-Cretaceous fossils possesses features that can link them uniquely to the Nymphaeales. Also, among the several herbaceous water plants reported from the Aptian–Albian of Brazil, none can be assigned unequivocally to the Nymphaeales²¹.

Phylogenetic analyses of the Nymphaeales based on combined molecular (*rbcl*, *matK*, and 18S rDNA) and non-molecular data resolve the order into two sister families: Cabombaceae (*Brasenia*, *Cabomba*) and Nymphaeaceae (*Nuphar*, *Barclaya*, *Ondinea*,

Nymphaea, *Euryale*, *Victoria*)²², as had been proposed on the basis of morphological characters alone²³. Among Nymphaeaceae, *Nuphar* is the sister group to the remaining genera, which are defined as a group by the uniform occurrence of zonosulcate pollen. In this group, *Barclaya* is sister to the Nymphaeoidae (*Ondinea*, *Nymphaea*, *Euryale*, *Victoria*)²².

We assessed the phylogenetic position of the fossil flower in relation to genera of extant Nymphaeales by incorporating the fossil into a reduced morphological data set compiled for extant Nymphaeales²², using those floral and pollen characters that were available for the fossil flower. Parsimony analysis supports the inclusion of the fossil in the Nymphaeales as the sister taxon to all extant Nymphaeaceae. Synapomorphies that link the fossil flower to Nymphaeaceae but separate it from Cabombaceae are the syncarpous gynoecium, the perigynous/epigynous perianth and androecium (hypogynous in *Nuphar*), and the presence of a central protrusion of the floral apex between the carpels (absent in *Nuphar* and *Barclaya*). Features acquired by some or all Nymphaeaceae, but that are not present in the fossil flower, include zonosulcate pollen, a perianth with four or more sepals, stamens in a spiral arrangement and with laminar filaments, and a continuous stigmatic ring in the centre of the flower.

Apparently plesiomorphic characters for the Nymphaeales that are retained both by the fossil and Cabombaceae are the trimerous perianth (six perianth parts in two whorls of three in Cabombaceae, six perianth parts in the fossil), monocolpate pollen, floral parts in apparent whorls, linear stamen filaments, and the separate radiating stigmatic areas. Two of these features (monocolpate pollen and discontinuous stigmatic area) are also retained in the earliest diverging genus of extant Nymphaeaceae (*Nuphar*).

The only feature of the fossil not seen in extant Nymphaeales is the reticulate tectum of the pollen wall, but this character occurs in all ANITA clades (except Amborellaceae), as well as in other key magnoliid groups such as Chloranthaceae. It is also present in many basal members of the monocots and eudicots and is probably the plesiomorphic condition for angiosperms as a whole. In light of the fossil material, a pollen wall with a closed tectum probably evolved independently in Cabombaceae and Nymphaeaceae.

Insect pollination has been inferred for most of the early angiosperms described to date on the basis of features such as abundant connective tissue, ethereal oil cells and valvate anther dehiscence in stamens of Early Cretaceous flowers, as well as the presence in the

Portuguese floras of insect coprolites consisting exclusively of angiosperm pollen¹⁰. Insect pollination is predominant in the Nymphaeales and also seems probable for the fossil water lily described here.

With the exception of species that are self-pollinated and cleistogamous (*Euryale*, 1 sp.; *Barclaya*, 4 spp., possibly also fly pollinated), or pollinated by *Trigona* bees (*Ondinea*, 1 sp.), most Nymphaeaceae (*Nuphar*, 7–25 spp.; *Nymphaea*, ~50 spp.; *Victoria*, 2 spp.) have large flowers comprising numerous parts (up to 70 petals, 750 stamens, 40 carpels) and are beetle pollinated²⁴. In contrast, flowers of Cabombaceae are small with few parts, and are either wind pollinated (*Brasenia*, 1 sp.) or pollinated by flies (*Cabomba*, 5 spp.)²³. The fossil flower described here is structurally most similar to smaller, simpler flowers of Cabombaceae than to the more complex flowers of extant Nymphaeaceae, and its early occurrence in the fossil record is consistent with the interpretation that large, multiparted, beetle-pollinated flowers are a secondary evolutionary development in Nymphaeales. This contrasts with traditional interpretations of angiosperm floral evolution that have emphasized pervasive patterns of reduction^{25,26}, but parallels the secondary increases in the number of floral organs that have been proposed for other basal angiosperms²⁷.

A survey of the Vale de Agua fossil flora, and of four other Barremian–Aptian assemblages from Portugal (Buarcos, Catefica, Famalicão, Torres Vedras), has documented the presence of 140–150 different angiosperm taxa represented by flowers, fruits and seeds, of which most (85%) are either plants at the magnoliid grade or monocots¹⁰. Until now, only a single one of the basalmost families (Chloranthaceae) had been identified, on the basis of flowers, fruits and dispersed stamens, in the Early Cretaceous²⁸. The discovery of Nymphaeales in this assemblage extends the fossil record of another extant angiosperm lineage into the oldest sediments known to contain angiosperm flowers, and is also consistent with the early diverging position of this group in molecular phylogenetic analyses. However, the large number of other taxa that cannot be assigned to extant magnoliid families implies significant extinction at the magnoliid grade and raises important issues about the level of sampling of total magnoliid diversity provided by extant taxa.

The diversity of the Portuguese flora based on mesofossils also stands in marked contrast to the paucity of other remains of angiosperms from the same strata, and from strata of similar age in other regions of the world. In contemporaneous dispersed pollen

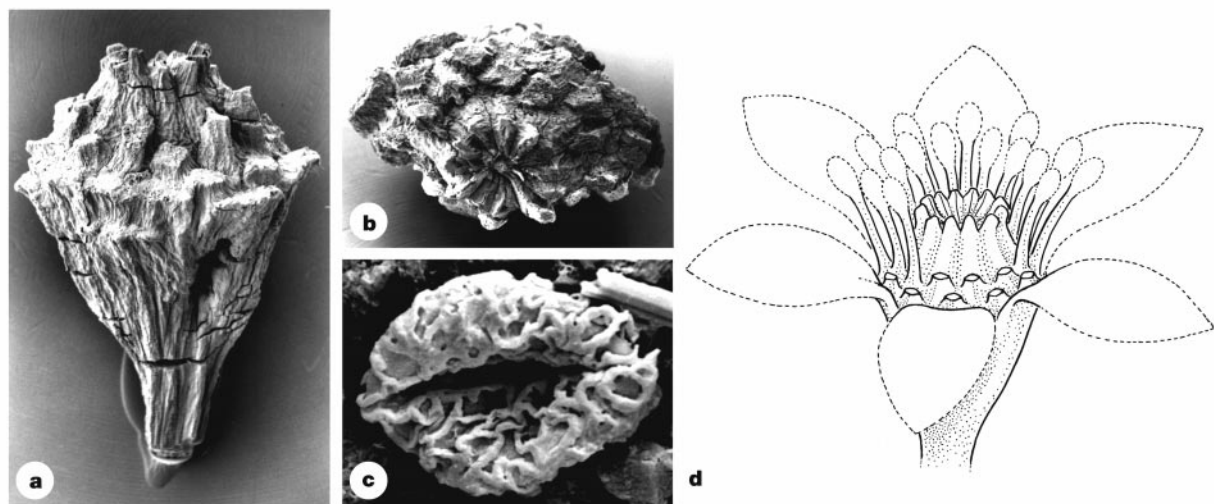


Figure 1 Fossil water lily flower from the Early Cretaceous (Barremian or Aptian) of Portugal, **a–c**, SEM micrographs of specimen S122015, Vale de Agua sample 328. **a**, Lateral view of fossil showing the perigynous nature of the flower; magnification, $\times 18$. **b**, Apical view of flower showing carpels radiating from the central protrusion of the floral

apex, and remnants from stamens (rhombic) and perianth (narrow elliptical); magnification, $\times 20$. **c**, Pollen grain showing monocolpate aperture and reticulum of pollen wall; magnification, $\times 2,000$. **d**, Reconstruction of the fossil flower.

floras, angiosperms typically constitute only a very small percentage of the total diversity^{15,17,29}—perhaps reflecting low pollen production and poor dispersal abilities associated with insect pollination. Similarly, with one strongly disputed exception angiosperm wood has not been recorded from Aptian or older rocks, and angiosperm leaves in Aptian or earlier floras are also extremely rare. However, exceptionally preserved whole plants reported from the Lower Cretaceous Crato Formation, Brazil, document that diverse herbaceous water plants were present by the Aptian–Albian and were a prominent part of the angiosperm assemblage of this flora²¹. These observations suggest that the apparent discrepancy between the diversity of angiosperm reproductive structures and the diversity of leaves and wood during the earliest phases of angiosperm diversification may in part be explained by the low potential of leaves and stems of herbaceous plants, including water lilies and monocots, to be preserved.

Received 20 October; accepted 15 December 2000.

- Qiu, Y.-L. *et al.* The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* **402**, 404–407 (1999).
- Qiu, Y.-L. *et al.* Phylogeny of basal angiosperms: analyses of five genes from three genomes. *Int. J. Plant Sci.* **161** (Suppl. 6), S3–S27 (2000).
- Soltis, P. S., Soltis, D. E. & Chase, M. W. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* **402**, 402–404 (1999).
- Kuzoff, R. A. & Gasser, C. S. Recent progress in reconstructing angiosperm phylogeny. *Trends Plant Sci.* **5**, 330–336 (2000).
- Friis, E. M., Pedersen, K. R. & Crane, P. R. Angiosperm floral structures from the Early Cretaceous of Portugal. *Pl. Syst. Evol.* **8** (Suppl.), 31–49 (1994).
- Magallón, S., Crane, P. R. & Herendeen, P. S. Phylogenetic pattern, diversity and diversification of eudicots. *Ann. Missouri Bot. Gard.* **86**, 297–372 (1999).
- Friis, E. M., Friis, E. M. Magnoliid reproductive organs from the Cenomanian–Turonian of north-western Kazakhstan: Magnoliaceae and Illiciaceae. *Plant Syst. Evol.* **216**, 265–288 (1999).
- Gandolfo, M. A., Nixon, K. C. & Crepet, W. L. in *Monocots: Systematics and Evolution* (eds Wilson, K. I. & Morrison, D. A.) 44–51 (CSIRO, Melbourne, 2000).
- Friis, E. M., Pedersen, K. R. & Crane, P. R. Early angiosperm diversification: the diversity of pollen associated with angiosperm reproductive structures in Early Cretaceous floras from Portugal. *Ann. Missouri Bot. Gard.* **86**, 259–296 (1999).
- Friis, E. M., Pedersen, K. R. & Crane, P. R. Reproductive structure and organization of basal angiosperms from the Early Cretaceous (Barremian or Aptian) of Western Portugal. *Int. J. Plant Sci.* **161** (Suppl. 6), S169–S182 (2000).
- Zbyszewski, G., Manupella, G. & Da Veiga Ferreira, O. *Carta geológica de Portugal na escala de 1/50 000. Notícia explicativa da folha 27-A Vila Nova de Ourém* (Serviços Geológicos de Portugal, Lisbon, 1974).
- Doyle, J. A. & Hickey, L. J. in *Origin and Early Evolution of Angiosperms* (ed. Beck, C. B.) 139–206 (Columbia Univ. Press, New York, 1976).
- Penny, J. H. J. An Early Cretaceous angiosperm pollen assemblage from Egypt. *Special Papers Palaeontol.* **35**, 121–134 (1986).
- Doyle, J. A. Revised palynological correlations of the lower Potomac Group (USA) and the Cocobeach sequence of Gabon (Barremian–Aptian). *Cretaceous Res.* **13**, 337–349 (1992).
- Hughes, N. F. & McDougall, A. B. Barremian–Aptian angiosperm pollen records from southern England. *Rev. Palaeobot. Palynol.* **65**, 145–151 (1990).
- Doyle, J. A. & Robbins, E. I. Angiosperm pollen zonation of the continental Cretaceous of the Atlantic Coastal Plain and its application to deep wells in the Salisbury Embayment. *Palynology* **1**, 43–78 (1977).
- Hughes, N. F. *The Enigma of Angiosperm Origins* (Cambridge Univ. Press, Cambridge, 1994).
- Rey, J. Recherches géologiques sur le Crétacé inférieur de l'Estremadura (Portugal). *Serviços Geológicos de Portugal, Memórias (Nova Série)* **3** 21, 1–477 (1972).
- Endress, P. K. & Igersheim, A. Gynoecium structure and evolution in basal angiosperms. *Int. J. Plant Sci.* **161** (Suppl. 6), S211–S223 (2000).
- Saporta, G. D. *Flore fossile du Portugal. Nouvelles contributions à la flore Mésozoïque. Accompagnées d'une notice stratigraphique par Paul Choffat* (Imprimerie de l'Académie Royale des Sciences, Lisbon, 1894).
- Mohr, B. & Friis, E. M. Early angiosperms from the Aptian Crato Formation (Brazil), a preliminary report. *Int. J. Plant Sci.* **161** (Suppl. 6), S155–S167 (2000).
- Les, D. H. *et al.* Phylogeny, classification and floral evolution of water lilies (Nymphaeaceae; Nymphaeales): A synthesis of non-molecular, *rbcl*, *matK*, and rDNA data. *Syst. Bot.* **24**, 28–46 (1999).
- Williamson, P. S. & Schneider, E. I. Cabombaceae. in *The Families and Genera of Vascular Plants. II Flowering plants—Dicotyledons. Magnoliid, Hamamelid and Caryophyllid Families* (eds Kubitzki, K., Rohwer, J. G. & Bittrich, V.) 157–161 (Springer, Berlin, 1993).
- Schneider, E. I. & Williamson, P. S. Nymphaeaceae. in *The Families and Genera of Vascular Plants. II Flowering plants—Dicotyledons. Magnoliid, Hamamelid and Caryophyllid Families* (eds Kubitzki, K., Rohwer, J. G. & Bittrich, V.) 486–493 (Springer, Berlin, 1993).
- Bessey, C. E. The phylogenetic taxonomy of flowering plants. *Ann. Missouri Bot. Gard.* **2**, 109–164 (1915).
- Takhtajan, A. *Flowering Plants. Origin and Dispersal* (Oliver & Boyd, Edinburgh, 1969).
- Drinnan, A. N., Crane, P. R. & Hoot, S. B. Patterns of floral evolution in the early diversification of non-magnoliid dicotyledons (eudicots). *Plant Syst. Evol.* **8** (Suppl.), 93–122 (1994).
- Friis, E. M., Crane, P. R. & Pedersen, K. R. in *Evolution and Diversification of Land Plants* (eds Iwatsuki, K. & Raven, P. H.) 121–156 (Springer, Tokyo, 1997).

29. Brenner, G. J. & Bickoff, I. S. Palynology and the age of the Lower Cretaceous basal Kurnub Group from the coastal plain to the northern Negev of Israel. *Palynology* **16**, 137–185 (1992).

Acknowledgements

We thank P. K. Endress and J. Schönenberger for valuable comments and help; and P. von Knorring for preparing the reconstruction of the fossil flower. The work was supported by grants from the Swedish Natural Science Foundation (to E.M.F.), the Carlsberg Foundation (to K.R.P. and E.M.F.), the Danish Natural Science Research Council (to K.R.P.) and the US National Science Foundation (to P.R.C.).

Correspondence and requests for materials should be addressed to E.M. Friis (e-mail: else.marie.friis@nrm.se).

Mesoscale vertical motion and the size structure of phytoplankton in the ocean

Jaime Rodríguez*, Joaquín Tintoré†, John T. Allen‡, José M^a Blanco*, Damià Gomis†, Andreas Reul*, Javier Ruiz§, Valeriano Rodríguez*, Fidel Echevarría§ & Francisco Jiménez-Gómez*||

* Departamento de Ecología, Universidad de Málaga, Campus de Teatinos, 29071-Málaga, Spain

† Institut Mediterrani d'Estudis Avançats (CSIC-UIB), 07071 Palma de Mallorca, Spain

‡ Southampton Oceanography Center, J. Rennell Division, Southampton SO14 3ZH, UK

§ Departamento de Biología Animal, Biología Vegetal y Ecología, Facultad de Ciencias del Mar, Universidad de Cádiz, 11510 Puerto Real, Cádiz, Spain

Phytoplankton size structure is acknowledged as a fundamental property determining energy flow through 'microbial' or 'herbivore' pathways¹. The balance between these two pathways determines the ability of the ecosystem to recycle carbon within the upper layer or to export it to the ocean interior¹. Small cells are usually characteristic of oligotrophic, stratified ocean waters, in which regenerated ammonium is the only available form of inorganic nitrogen and recycling dominates. Large cells seem to characterize phytoplankton in which inputs of nitrate enter the euphotic layer and exported production is higher^{2–4}. But the size structure of phytoplankton may depend more directly on hydrodynamical forces than on the source of available nitrogen^{5–7}. Here we present an empirical model that relates the magnitude of mesoscale vertical motion to the slope of the size–abundance spectrum^{8–10} of phytoplankton in a frontal ecosystem. Our model indicates that the relative proportion of large cells increases with the magnitude of the upward velocity. This suggests that mesoscale vertical motion—a ubiquitous feature of eddies and unstable fronts—controls directly the size structure of phytoplankton in the ocean.

The oceanic mesoscale, 10–100 km, is the equivalent of the atmospheric storm scale. It is generally considered to be the most energetic scale, and it is where fronts between water masses become unstable and strong three-dimensional circulations are set up. Fronts are places of enhanced biological activity and, from the biological point of view, vertical circulation is of great significance to primary productivity as it may explain the patchiness of primary productivity in the surface layers of the ocean and the patchiness of nutrient distributions¹¹. During the OMEGA (Observations and

|| Present address: Departamento de Biología Animal, Biología Vegetal y Ecología, Facultad de Ciencias Experimentales, Universidad de Jaén, 23071 Jaén, Spain.

Modelling of Eddy Scale Geostrophic and Ageostrophic Circulation) project, multidisciplinary observations were made in the northwestern Alborán (Mediterranean) Sea, where water of Atlantic and Mediterranean origin meet, creating gyres, fronts and eddies of different scales. This sea is an ideal region in which to study physical–biological coupling, particularly the relation between vertical motion and the size structure of phytoplankton.

If vertical motion controls the size structure of phytoplankton, the slope of the community size–abundance spectrum (SAS; see Methods) should be less negative in areas of upward motion than in areas of downward motion, because upward motion increases the residence time of large cells in the upper layer against their tendency to sink. Small cells, on the other hand, do not sink (or sink very slowly) because of the dominance of viscosity at very low Reynolds numbers; therefore, vertical motion would not have a differential effect on this part of the community SAS. Ideally there should be a positive correlation between the magnitude of the upward vertical velocity (w) and the flatness of the spectrum linear model, owing to the increased retention of large cells within the upper layers. The relation obtained for our complete set of observations was not so simple (Fig. 1a); however, a linear relationship between vertical

velocity and size spectrum slope was clearly evident for vertical velocity magnitudes of less than 5–6 m d^{-1} (Fig. 1b):

$$\text{SAS slope} = -0.944 + 0.033w \text{ (in } \text{m d}^{-1}\text{)} \quad (1)$$

($r^2 = 0.683$, $\alpha < 0.001$; 95% confidence interval (CI) for the regression slope: 0.025–0.042). Our conceptual model indicates that there may be a positive relation between the magnitude of the mesoscale vertical motion and the slope of the SAS of phytoplankton. If the size structure of phytoplankton were nutrient dependent, we would expect a different relation between vertical velocity and SAS slope at the depth of the ‘deep chlorophyll maximum’ (DCM), where cells are expected to find optimal growth conditions in the vicinity of the nutricline.

This is not the case (Fig. 1b), and the relation is the same for assemblages at, over, or below the DCM. This suggests that nutrients are not the critical factor controlling the size structure of the community. We would expect, however, a light-related effect on the DCM size structure, as when the DCM occurs very deep in the water column it seems to be characterized by small cells and low amounts of chlorophyll, whereas the opposite occurs in the case of a shallow DCM^{4,12,13}. Our observations do not support this hypothesis (Fig. 2). In fact, the size spectrum with the least negative slope corresponds to a very deep DCM community (Fig. 3), whereas spectra very similar in shape (Fig. 2) correspond to DCM layers placed at very different depths within the water column (Fig. 3). In all the cases, size distributions are highly similar up to cell sizes around 5 μm equivalent spherical diameter and diverge for larger cells with higher sinking velocities. This divergence illustrates the size-dependent, counteracting effect of upward vertical motion and implies that, on relative terms, the flatter spectrum has around seven times the biomass of steeper spectra in the size range of cells between $10^4 \mu\text{m}^3$ (~25 μm diameter) and $10^5 \mu\text{m}^3$ (~60 μm diameter).

Hypotheses about the vertical motion control on the size of planktonic organisms were previously proposed at two very different scales. At the scale of tens of metres, which characterizes Langmuir circulation, the coupled upwelling and downwelling processes were considered to be responsible for the accumulation

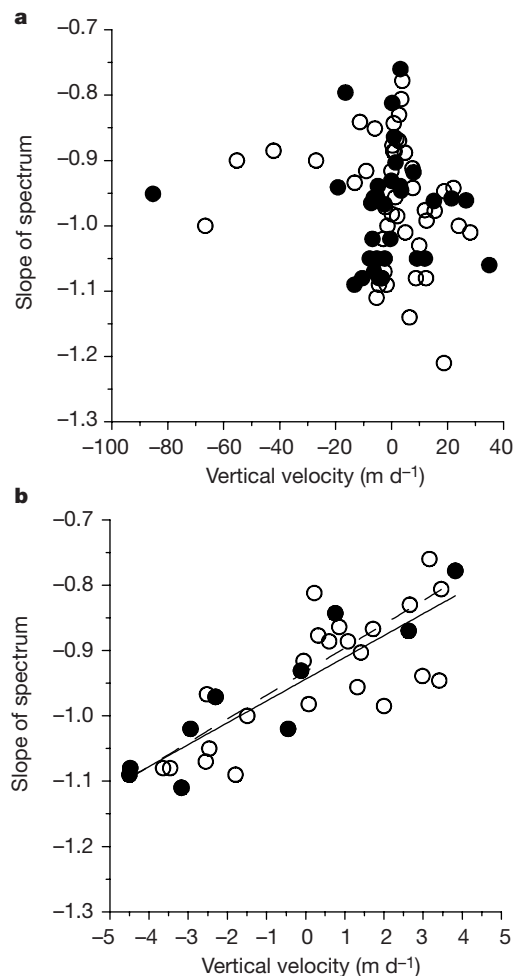


Figure 1 Slope of the size–abundance spectrum plotted against the magnitude of vertical velocity (w) diagnosed for mesoscale instability. **a**, For the full range of w , filled circles represent observations below the thermocline that can be considered free from the influence of mixed layer mixing processes. **b**, Data points and a linear regression (continuous line) for w in the range between $\pm 5 \text{ m d}^{-1}$; filled circles represent measurements at the depth of DCM, described by the model (broken line): SAS slope = $-0.932 + 0.036w$ ($r^2 = 0.854$; $\alpha < 0.001$; 95% CI for the regression slope = 0.024–0.048).

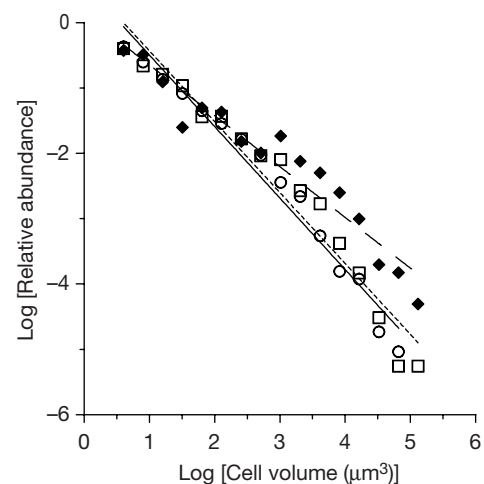


Figure 2 Size–abundance spectra of DCM phytoplankton communities at the spots of highest upward and downward vertical velocity in the range considered in Fig. 1b. Filled diamonds, DCM community placed at 60 m depth, $w = +3.8 \text{ m d}^{-1}$; open squares, 30 m depth, $w = -4.8 \text{ m d}^{-1}$; open circles, 50 m depth, $w = -4.8 \text{ m d}^{-1}$. Abundance is expressed in relative units to eliminate differences in absolute biomass between spectra. For filled diamonds, $r^2 = 0.934$ ($\alpha < 0.001$); regression slope = -0.778 (95% CI = $-0.897, -0.660$); for open squares, $r^2 = 0.952$ ($\alpha < 0.001$); regression slope = -1.085 (95% CI = $-1.224, -0.946$); for open circles, $r^2 = 0.972$ ($\alpha < 0.001$); regression slope = -1.095 (95% CI = $-1.207, -0.983$).

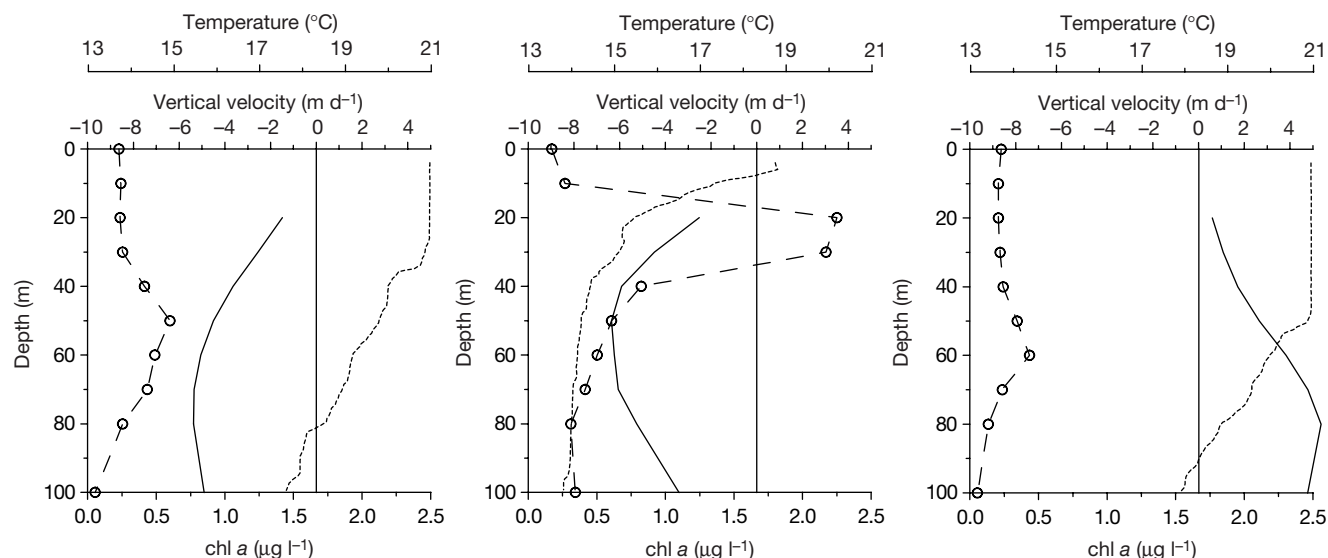


Figure 3 Vertical structure of the water column corresponding to size–abundance spectra in Fig. 2. Continuous line, vertical velocity; dotted line, temperature; open circles, chlorophyll (chl) concentration.

of organisms with negative or positive buoyancy (a size-related property)¹⁴. At the biogeographical, large scale of oceanic regions in the Pacific Ocean, it has been proposed that average cell diameter is the largest in regions of weak ascending and weak descending motion, and that it is the smallest in areas of most intensive vertical motion^{5,6}. Notably, a linear relation between cell diameter and vertical motion for the range of weak velocities at the large scale of oceanic convergences and divergences has been reported⁵.

Observations about the size distribution of fractionated chlorophyll in the equatorial Pacific are also better interpreted in terms of circulation (equatorial upwelling and horizontal transport) rather than a nutrient effect⁷. There have been no previous published studies relating the vertical component of mesoscale fluid flow to the size structure of the primary producer community for two main reasons. First, mesoscale vertical velocities are too small to be measured directly using traditional vessel mounted or moored acoustic current profilers and therefore they have to be determined (see Methods) through careful analysis of high-resolution hydrographic measurements and assumed balance conditions for the flow^{15–17}. Second, determining the SAS of the planktonic community (see Methods) is a time-consuming task frequently substituted by average descriptors such as “net-plankton, cell average diameter”^{5,6} or bulk variables such as “size-fractionated chlorophyll”⁷.

The impossibility of discriminating a nutrient-mediated effect of mesoscale vertical motion on the size structure of phytoplankton at the DCM layer gives support to the existence of a direct vertical motion control on community structure, and suggests that the interaction between water vertical motion and cell sinking properties are the primary forces in shaping the size–abundance spectrum of phytoplankton in the upper layer of the ocean. Biological processes such as photo-adaptation, nutrient uptake and growth, and cell motility would determine other DCM properties like species composition or chlorophyll and biomass concentration, usually analysed to explain the origin and maintenance of the DCM layer and the vertical distribution of phytoplankton in general^{18–21}.

However, we have to question why the interaction between vertical motion and size structure is limited to weak magnitude of vertical velocity. There are several, non-mutually exclusive explanations to be considered. The first is principally a sampling problem. Generally, the magnitude of vertical velocity is inversely related to the scale of the horizontal flow field^{16,22,23}. For example, vertical

velocities estimated from the curl of the wind stress^{5,6} are typically in the range of $\pm 10^{-1} \text{ m d}^{-1}$. In highly energetic mesoscale flows like ours, vertical velocities between 10 and 100 m d^{-1} (refs 16, 17, 24) are more typical, but are associated with strong advection (in our case up to 1.5 m s^{-1}) that make it difficult to make synoptic observations and result in vertical motion estimates that are susceptible to large errors²⁵. Consequently, it is out of the centre of the frontal jet, where the observations are more truly synoptic that we can have most confidence in our results and where the magnitude of the vertical velocity is lowest.

Second, mesoscale, ageostrophic vertical motion can be modified by other physical processes like wind-driven turbulence in the uppermost layer, tidal oscillations or internal waves at the thermocline that might contribute to the scatter of points at high vertical velocities. To avoid contamination with mixed layer processes, we have identified those data points below the thermocline (Fig. 1a). Both the complete data set and the subset containing data only from below the thermocline show similar scatters against vertical velocity magnitude, however, indicating that wind-driven turbulence is not the origin of the scatter. It is more difficult to avoid contamination from internal waves, but the SeaSoar hydrographic sections show no evidence of significant internal wave activity in the region.

All this suggests that the interaction of mesoscale vertical motion with other physical processes is probably of less importance in explaining the scatter of points at high values of vertical velocity. In the third place, the interaction between upward water motion and the sinking velocity of cells should be considered. The calculated range of vertical velocities includes the sinking velocities of phytoplanktonic cells even for the largest sizes¹⁴. The largest cells included in our size–abundance spectrum (Fig. 2) have diameters between 60 and $70 \mu\text{m}$ and, according to Smayda’s model¹⁴, they would sink at $2\text{--}2.5 \text{ m d}^{-1}$. Thus, the range of vertical velocities in Fig. 1b is of the same order as the range of the size-dependent sinking velocity of the cells included in our SAS analysis. This suggests that only a relatively small upward water motion is needed to support phytoplanktonic cells in the size range analysed, which certainly contains the bulk of ocean primary producers. Consequently, an extended mathematical model for the whole range of vertical velocity values shown in Fig. 1 would need a much broader SAS analysis including large particles such as colonies, aggregates and fecal pellets. Such an analysis would permit the complete evaluation of the role of mesoscale dynamics in controlling the size structure of pelagic particles and its impact on the vertical flux of carbon in the ocean. □

Methods

Modelling the SAS of phytoplankton

We counted and measured phytoplanktonic cells by using a video-interactive, microscopy-image analysis system that permits the discrimination and elimination of detritus and other non-living particles. Individual linear measurements were transformed into estimations of cell volume through numerical computation, and these were then distributed along an octave ($\log_2$) scale of cell volume. Finally, a regression mode 1 was applied to the log-transformed values of cell abundance (y axis, cells per ml) and cell volume (x axis, μm^3). The resulting key parameter is the slope (b) of the regression model:

$$\log[\text{cell abundance (cells per ml)}] = a - b \log[\text{cell volume } (\mu\text{m}^3)]$$

A general account of methods for the analysis of the size structure of planktonic communities can be found in ref. 10.

Computation of vertical velocity

In the framework of the Quasi-Geostrophic (QG) approximation, the vertical component of the velocity field (w) can be estimated from a diagnostic equation known as the Omega equation²⁶. This equation can be solved in a three-dimensional domain provided that (1) the QG vertical forcing is known at every point of the domain; and (2) boundary conditions are specified. The vertical forcing term can be computed by finite differences from hydrographic data alone (for example, from density and/or dynamic height data), which must have been previously interpolated from stations to a regular grid. Regarding boundary conditions, we assumed $w = 0$ at the bottom and surface, and Neumann conditions at the lateral boundaries²⁷. Several tests show that the sensitivity of vertical velocity fields to lateral boundaries restricts to the few outermost grid points, with no significant influence on the results obtained in the inner domain.

Received 6 September 2000; accepted 3 January 2001.

- Legendre, L. & Le Fèvre, J. in *Productivity of the Ocean: Present and Past* (eds Berger, W. H., Smetacek, V. S. & Wefer, G.) 49–63 (John Wiley & Sons Limited, New York, 1989).
- Eppley, R. W. & Peterson, B. J. Particulate organic matter flux and planktonic new production in the deep ocean. *Nature* **282**, 677–680 (1979).
- Malone, T. C. in *The Physiological Ecology of Phytoplankton* (ed. Morris, I.) 433–464 (Blackwell Scientific Publications, Oxford, 1980).
- Chisholm, S. W. in *Primary Productivity and Biogeochemical Cycles in the Sea* (eds Falkowski, P. G. & Woodhead, A. D.) 213–237 (Plenum, New York, 1992).
- Semina, H. J. Water movement and the size of phytoplankton cells. *Sarsia* **34**, 267–272 (1968).
- Semina, H. J. The size of phytoplankton cells in the Pacific Ocean. *Int. Revue ges. Hydrobiol.* **57**, 177–205 (1972).
- Peña, A., Lewis, M. R. & Harrison, G. Primary productivity and size structure of phytoplankton biomass on a transect of the equator at 135°W in the Pacific Ocean. *Deep-Sea Res.* **37**, 295–315 (1990).
- Platt, T. & Denman, K. H. The structure of pelagic marine ecosystems. *Rapp. P.-V. Reun. Cons. Int. Explor. Mer* **173**, 60–65 (1978).
- Rodríguez, J. & Mullin, M. M. Relation between biomass and body weight of plankton in a steady state oceanic ecosystem. *Limnol. Oceanogr.* **31**, 361–370 (1986).
- Rodríguez, J. & Li, W. K. W. (eds) The size structure and metabolism of the pelagic ecosystem. *Scientia Marina* **58**, 1–167 (1994).
- Strass, V. H. Chlorophyll patchiness caused by mesoscale upwelling at fronts. *Deep-Sea Res.* **39**, 75–96 (1992).
- Harris, G. P., Ganf, G. G. & Thomas, D. P. Productivity, growth rates and cell size distributions of phytoplankton in the SW Tasman Sea: implications for carbon metabolism in the photic zone. *J. Plankton Res.* **9**, 1003–1030 (1987).
- Rodríguez, J. *et al.* Patterns in the size structure of phytoplankton community in the deep fluorescence maximum of the Alboran Sea (southwestern Mediterranean). *Deep-Sea Res.* **45**, 1577–1593 (1998).
- Smayda, T. J. The suspension and sinking of phytoplankton in the sea. *Oceanogr. Mar. Biol. Annu. Rev.* **8**, 353–414 (1970).
- Leach, H. The diagnosis of synoptic-scale vertical motion in the seasonal thermocline. *Deep-Sea Res.* **34**, 2005–2017 (1987).
- Tintoré, J., Gomis, D., Alonso, S. & Parrilla, G. Mesoscale dynamics and vertical motion in the Alboran Sea. *J. Phys. Oceanogr.* **21**, 811–823 (1991).
- Allen, J. T. & Smeed, D. A. Potential vorticity and vertical velocity at the Iceland Faroes front. *J. Phys. Oceanogr.* **26**, 2611–2634 (1996).
- Cullen, J. J. The deep chlorophyll maximum: comparing vertical profiles of chlorophyll *a*. *Can. J. Fish. Aquat. Sci.* **39**, 791–803 (1982).
- Furuya, K. & Marumo, R. The structure of the phytoplankton community in the subsurface chlorophyll maxima in the western North Pacific Ocean. *J. Plankton Res.* **5**, 393–406 (1983).
- Estrada, M. *et al.* Variability of deep chlorophyll maximum characteristics in the Northwestern Mediterranean. *Mar. Ecol. Prog. Ser.* **92**, 289–300 (1993).
- Ruiz, J., García, C. M. & Rodríguez, J. Vertical patterns of phytoplankton size distribution in the Cantabric and Balearic seas. *J. Mar. Systems* **9**, 269–282 (1996).
- Fiekas, V., Leach, H., Mirbach, K.-J. & Woods, J. D. Mesoscale instability and upwelling. Part I: Observations at the North Atlantic intergyre front. *J. Phys. Oceanogr.* **24**, 1750–1758 (1994).
- Viúdez, A., Tintoré, J. & Haney, R. L. Circulation in the Alboran Sea as determined by quasi-synoptic hydrographic observations. I. Three dimensional structure of the two anticyclonic gyres. *J. Phys. Oceanogr.* **26**, 684–705 (1996).
- Pollard, R. T. & Regier, L. Large variations in potential vorticity at small spatial scales in the upper ocean. *Nature* **348**, 227–229 (1990).
- Allen, J. T., Smeed, D. A., Nurser, A. J. G., Zhang, J. W. & Rixen, M. Diagnosing vertical velocities with the QG omega equation: an examination of the errors due to sampling strategy. *Deep-Sea Res.* (in the press).

- Cushman-Roisin, B. *Introduction to Geophysical Fluid Dynamics* (Prentice-Hall, Englewood Cliffs, New Jersey, 1994).
- Gomis, D., Ruiz, S. & Pedder, M. A. Diagnostic analysis of the 3D ageostrophic circulation from a Multivariate Spatial Interpolation of CTD and ADCP data. *Deep-Sea Res.* **48**, 269–295 (2000).

Acknowledgements

This work was supported by the MAST III programme of the European Commission and the CICYT-CYTMAR programme (Spain). We thank M. Emelianov for translating the Russian work of H. J. Semina. We also thank the officers, technicians and crew of *BIO Hesperides* for their help during the OMEGA cruise.

Correspondence and requests for materials should be addressed to J.R. (e-mail: jaime@uma.es).

Ecological importance of trichromatic vision to primates

Nathaniel J. Dominy & Peter W. Lucas

Department of Anatomy, University of Hong Kong, 5 Sassoon Road, Hong Kong, People's Republic of China

Trichromatic colour vision, characterized by three retinal photopigments tuned to peak wavelengths of ~430 nm, ~535 nm and ~562 nm (refs 1, 2), has evolved convergently in catarrhine primates and one genus of New World monkey, the howlers (genus *Alouatta*)³. This uniform capacity to discriminate red-green colours, which is not found in other mammals, has been proposed as advantageous for the long-range detection of either ripe fruits^{4,5} or young leaves⁶ (which frequently flush red in the tropics⁷) against a background of mature foliage^{8,9}. Here we show that four trichromatic primate species in Kibale Forest, Uganda, eat leaves that are colour discriminated only by red-greenness, a colour axis correlated with high protein levels and low toughness. Despite their divergent digestive systems, these primates have no significant interspecific differences in leaf colour selection. In contrast, eaten fruits were generally discriminated from mature leaves on both red-green and yellow-blue channels and also by their luminance, with a significant difference between chimpanzees and monkeys in fruit colour choice. Our results implicate leaf consumption, a critical food resource when fruit is scarce¹⁰, as having unique value in maintaining trichromacy in catarrhines.

In Kibale National Park, Uganda, leaves eaten by trichromatic chimpanzees (*Pan troglodytes*), black-and-white colobus (*Colobus guereza*), red colobus (*Piliocolobus badius*) and red-tailed monkeys (*Cercopithecus ascanius*) could be discriminated from mature leaves in colour only by the red-green channel (Fig. 1a). Consumed leaves, which included some leaves of mature status, had a higher average luminance, but the red-green colour channel was a far better discriminant (Fig. 1b), primarily because many young red leaves were dark. There were no significant differences overall between primate species in the colours of leaves eaten (Duncan's multiple range test, $P > 0.05$), and the red-greenness of the leaves was highly positively correlated with the ratio of protein content to toughness (Spearman's test, $R = 0.58$, $P < 0.001$). This ratio was consistently higher in leaves consumed by each primate than in mature foliage (Fig. 2).

In contrast, eaten fruits were distinguishable on both the red-green and yellow-blue colour channels (Fig. 3), as well as luminance (Duncan's multiple range test, $P < 0.001$ for each case). Fruits selected by monkeys were more yellow than mature foliage ($P < 0.03$), whereas those eaten by chimpanzees were significantly more yellow than those eaten by monkeys (Duncan's multiple range

test, $P < 0.001$), principally because these were fully ripe when chimpanzees ate them.

Many catarrhine primates and howler monkeys eat leaves at critical periods when fruits are unavailable, preferring 'young' leaves where possible¹¹. Our results support the value of trichromacy to a foraging catarrhine because the yellow–blue channel shared with dichromats fails to detect young leaves. Leaf age, however, is probably not a cue in itself; the link that we show between nutritional reward and leaf colour, the latter being more reliable than leaf size at detecting developmental state⁶ and salient at greater distance, offers a clear selective advantage. Although the colour of young leaves in temperate regions is nearly always green, a substantial proportion of those in tropical floras are red: 50–62% in the Old World and 18–36% in the New World^{7,12}. Redness results partly from 'delayed greening' where plants postpone chloroplast function until full leaf expansion^{7,13}. The anthocyanin pigments responsible for reddening may prevent photoinhibition, defend against fungi or deter invertebrate herbivores^{7,13}.

Regardless of function, the nutritional gain to a primate perceiving a red–green transition in leaves, to which a dichromat would be blind, lies in the ingestion of protein during developmental stages before mechanical toughening. The ratio of protein content to fibre (or protein to toughness, as toughness is the dominant sensory attribute of cell-wall fibre to which it is nonlinearly related¹⁴) is an important measure of leaf quality for primates¹⁵. As primate folivory is far less costly to plant fitness than that by invertebrates¹⁶, primates probably exert insufficient selective pressure for plants to modify leaf characteristics such as colour.

It is also unlikely that trichromatic primates influence fruit colour because there is no clear suite of fruit characteristics associated with them^{17–19}. This is partly a reflection of their competition against many other frugivores, like birds and bats, which are known to confer successful seed dispersal^{17,19}. It has been suggested that plants that seem to target primates as seed-dispersers produce large fruits with characteristic colours (such as yellow, orange or brown), a thick peel, single large seeds and acid-sweet flesh^{20,21}. If this hypothesis were true, however, it would not explain why the external properties of fruits seem to be consistent throughout the tropics, even where primates are principally dichromatic, as in the New World, or absent, as in New Guinea¹⁸.

A further weakness in suggesting co-adaptation between trichromacy and fruit colour is that one whole group of trichromatic primates, the colobine subfamily of Old World monkeys, are seed

destroyers, frequently consuming unripe fruits expressly for their immature seeds²². In addition, cercopithecines, the other subfamily of Old World monkeys, often select unripe fruits in Kibale, offering them a competitive advantage over apes although, very often, this is to the detriment of the seeds they process²³. Finally, the contention that primate trichromacy evolved to detect pre-existing fruit colours also seems implausible because, despite a high degree of frugivory and similar fruit colours to the Old World tropics¹⁸, most neotropical primates are basically dichromatic.

The fruit-feeding hypothesis for trichromacy in higher primates is unconvincing. In this study, fruits selected by sympatric primates not only differed significantly in colour, but also the yellow–blue channel, common to all dichromatic mammals, could distinguish

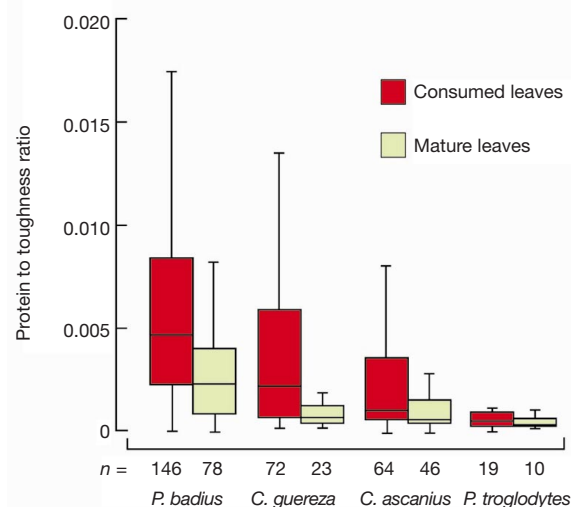


Figure 2 Protein (in %BSA equivalents) to toughness (in J m^{-2}) ratios of leaves eaten by the primates compared with those of mature leaves present on the same trees (Duncan's multiple range test, $P < 0.001$). Protein alone was higher in eaten leaves, whereas toughness was lower. Use of a ratio follows analogous previous use¹⁵. Tannins interfere with protein uptake, but their levels in leaves in this study were relatively low, which is well known in Kibale³⁰. Plots show medians; boxes enclose 50% of data; bars show ranges. A higher protein : toughness ratio reflects greater ease in opening leaf cells to release more nutritious cellular contents. There was no difference in the red–greenness of leaves between primates.

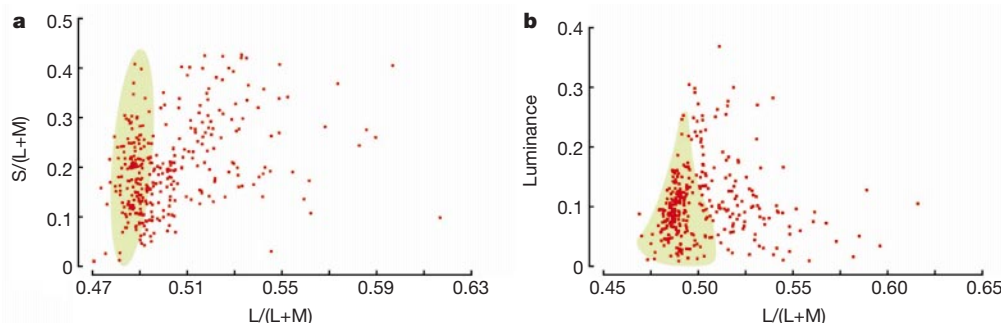


Figure 1 Distribution of leaf colours eaten by the four primates (red squares) in relation to the domain of mature foliage (outlined in green). **a**, Eaten leaves are distinguished on the horizontal $L/(L+M)$ 'red–green' axis ($F = 78.1$, $P < 0.001$), where increasing redness lies to the right side, but not on the vertical $S/(L+M)$ 'yellow–blue' axis ($F = 0.347$, $P = 0.556$). **b**, Differences in red–greenness are far greater than differences in luminance (plotted as a proportion of reflectance from a compacted barium-sulphate powder standard), as shown by the variance ratio for the latter ($F = 13.27$, $P < 0.001$). Some mature leaves were eaten, but others simply had a mature colour, differing in other physico-chemical properties. The proportion of feeding time spent consuming

non-mature leaves was 61.6, 66.7, 32.4 and 4.4% for *Piliocolobus badius*, *Colobus guereza*, *Cercopithecus ascanius* and *Pan troglodytes*, respectively. Leaf eating by chimpanzees is low at Kanyawara owing to frequent consumption of terrestrial herbaceous vegetation. Elsewhere, leaf consumption by chimpanzees is much more common²⁹. Consumption of mature leaves (that is, of the background itself) was sporadic: 3.8% in *P. badius*, 14.8% in *C. guereza*, 1.5% in *C. ascanius*, and 1.0% in *P. troglodytes*. In some key plant species, leaf flushing was year-round, but there were two re-leafing peaks that seemed to coincide with the onset of the wet season.

many fruits from leaves (Fig. 3). This would explain why dichromatic tropical mammalian frugivores, such as all New World primates except *Alouatta*, can find fruits of the same colour range with no apparent difficulty¹⁹. Indeed, howler monkeys are sluggish primates in comparison with sympatric New World monkeys, such as spider monkeys that feed on the same general range of fruit species, and are rarely likely to be first into fruiting trees.

We thus argue that our results support the value of trichromacy to a primate foraging for leaves. In all four species studied, the overall uniformity of leaf colour choice matches the uniformity in their photopigment tuning. This is despite the two colobine species, with ruminant stomachs, having diets of differing plant composition in comparison with those of the simple-stomached red-tail monkey and chimpanzee. The potential nutritional reward in detecting the red–green transition would thus be valuable to all these primates regardless of digestive physiology (Fig. 2). To an individual dichromatic primate, a young red leaf might appear dark and mature and thus tough and unpalatable⁶.

Selection pressures that maintain trichromacy in primates are clearly indicated by the near absence of individual variation outside *Homo sapiens* (such as the genus *Macaca*, where <0.1% of individuals have been found to be dichromatic²⁴). Because leafing and fruiting are distinct alternating events both in Kibale and elsewhere in Africa¹⁰, the regular biannual reliance of catarrhine primates on young leaf matter may account for the selective pressures resulting in uniform trichromacy, regardless of whether these primates are effective seed-dispersing frugivores or not. Few diurnal tropical mammals feed on a generalized range of leaves in the manner of higher primates. Of New World primates, only *Alouatta* spp. and *Brachyteles arachnoides* (whose retinal cone pigments have not been examined) feed on leaves to any great extent. The contrast between this and the dependence of most catarrhines on leaves, either as a critical resource when there is a fruit dearth or as in colobines as a staple, has been explained by differing inter-continental patterns of leaf phenology¹⁰.

Competition for ripe fruit attracts many other frugivores that signal its availability. Although the red–green colour channel can help a group to identify such fruits at long range⁸, the yellow–blue channel can often do this too⁹. Furthermore, many closer-range pre-ingestive cues, including smell and texture, reinforce visual indications. In contrast, finding palatable leaves is, on any leaf patch with leaves of differing states, more purely an individual task for which the value of the red–green channel is very clear. Its sensitivity is paramount in that very small colour changes in leaves reflect a large change in their quality (Figs 1 and 2). We believe that full trichro-

matic vision evolved originally for leaf foraging in higher primates. Its independent evolution in folivorous howler monkeys in the neotropics, where proposed fruiting/re-leafing synchrony¹⁰ and a possible lower incidence of red leaves may have decreased the overall importance of trichromacy to most primate species, supports this contention. □

Methods

Animals and study area

Studies were carried out at the Makerere University Biological Field Station (Kanyawara) in Kibale National Park, Uganda (0° 13' to 0° 41' N; 30° 19' to 30° to 32' E). The site is moist evergreen rainforest²⁵. The primates were observed for 1,170 h between January and October 1999. Focal animal techniques with several observers were used to maximize data. New focal animals were selected every 10 min. We collected a total of 386 h of feeding observations on these primates, in which time they ate parts of 118 plant species.

Food samples

Specimens of all observable food items were systematically collected and analysed for colour, chemical content and toughness. All foods were processed fresh with the aid of a field kit designed to assess physico-chemical food characteristics²⁶. Food samples were either those dropped by primates or, if not, vacated trees were climbed to collect food items with fresh bite marks. Full physico-chemical characteristics were obtained for 350 eaten leaves, 164 mature leaves and 110 eaten fruits, which formed the dataset analysed here.

Physico-chemical methods

An Ocean Optics S2000 spectrometer was connected to, and powered by, a laptop PC through a DAQCard1200 PCMCIA card. Illumination was provided by an Ocean Optics LS-1 12-volt 3100 K tungsten halogen lamp giving adequate light between 350 and 700 nm. For colour recording, small (~3 mm × 3 mm) food samples were placed in a purpose-built chamber with both light conduction and transmission to the spectrometer via 200–400 µm Ocean Optics patch fibre optic cables equipped with lenses. The food reflectance spectra were decoded into potential excitations of short (S), medium (M) and long (L) wavelength retinal cone pigments, following which the two colour channels of primate trichromatic vision were constructed²⁷: yellow–blue, S/(L+M); and red–green, L/(L+M). For protein analysis (spectrophotometry), cables were connected to an Ocean Optics cuvette holder for 1-cm quartz cuvettes. Results for the Coomassie blue (Coomassie brilliant blue G-250; Sigma) test for protein were expressed as a percentage of bovine serum albumin (%BSA) equivalents. Radial diffusion assays were run for tannins, results being expressed as percentage quebracho equivalents. We measured toughness with a portable universal testing machine²⁸. Data were analysed by analysis of variance after logarithmic transformation of the variables (except yellow–blue signals, which were normally distributed; Fig. 3).

Received 7 September; accepted 22 December 2000.

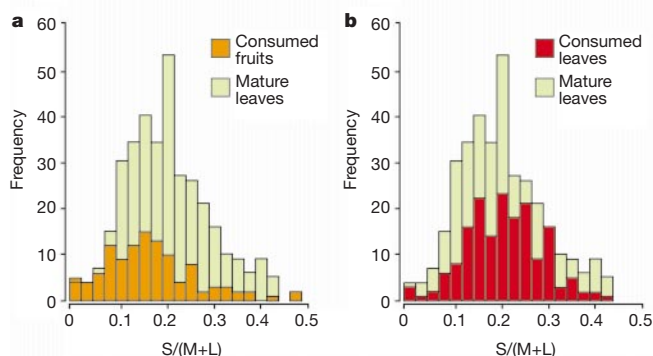


Figure 3 Histograms of colour distributions on the S/(L+M) 'yellow–blue' axis of eaten leaves and fruits in comparison with mature leaves. **a**, The distributions of eaten fruits and mature leaves overlap, but mainly the peak of the fruit distribution lies significantly to the left ($F = 12.7$, $P < 0.001$), reflecting the more yellow hue of many fruits. **b**, There is no difference on the yellow–blue channel between the distribution of eaten leaves and mature foliage ($F = 0.347$, $P = 0.556$).

- Li, W.-H., Tan, Y., Boissinot, S., Shyue, S. K. & Hewett-Emmett, D. in *The Biology of Biodiversity* (ed. Kato, M.) 259–274 (Springer, Tokyo, 2000).
- Jacobs, G. H. & Deegan, J. F. II Uniformity of color vision in Old World monkeys. *Proc. R. Soc. Lond. B* **266**, 2023–2028 (1999).
- Kainz, P. M., Neitz, J. & Neitz, M. Recent evolution of trichromacy in a New World monkey. *Vis. Res.* **38**, 3315–3320 (1998).
- Allen, G. *The Colour Sense: Its Origin and Development* (Trübner, London, 1879).
- Regan, B. C. et al. Frugivory and color vision in *Alouatta seniculus*, a trichromatic platyrrhine monkey. *Vis. Res.* **38**, 3321–3327 (1998).
- Lucas, P. W., Darvell, B. W., Lee, P. K. D., Yuen, T. D. B. & Choong, M. F. Colour cues for leaf food selection by long-tailed macaques (*Macaca fascicularis*) with a new suggestion for the evolution of trichromatic color vision. *Folia Primatol.* **69**, 139–152 (1998).
- Coley, P. D. & Kursar, T. A. in *Tropical Forest Plant Ecology* (eds Mulkey, S. S., Chazdon, R. L. & Smith, A. P.) 305–336 (Chapman & Hall, New York, 1996).
- Sumner, P. & Mollon, J. D. Catarrhine photopigments are optimized for detecting targets against a foliage background. *J. Exp. Biol.* **203**, 1963–1986 (2000).
- Sumner, P. & Mollon, J. D. Chromaticity as a signal of ripeness in fruits taken by primates. *J. Exp. Biol.* **203**, 1987–2000 (2000).
- Terborgh, J. & van Schaik, C. P. in *Organization of Communities Past and Present* (eds Gee, J. H. R. & Giller, P. S.) 205–226 (Blackwell Scientific, Oxford, 1987).
- Clutton-Brock, T. H. *Primate Ecology* (Academic, London, 1977).
- Opler, P. A., Frankie, G. W. & Baker, H. G. Comparative phenological studies of treelet and shrub species in tropical wet and dry forests in the lowlands of Costa Rica. *J. Ecol.* **68**, 167–188 (1980).
- Juniper, B. E. in *International Dendrological Society Yearbook* 49–57 (International Dendrological Society, London, 1993).
- Lucas, P. W., Turner, I. M., Dominy, N. J. & Yamashita, N. Mechanical defences to herbivory. *Ann. Bot.* **86**, 913–920 (2000).
- Chapman, C. A. & Janson, C. H. in *Primate Communities* (eds Fleagle, J. G., Janson, C. & Reed, K. E.) 237–267 (Cambridge Univ. Press, Cambridge, 1999).
- Coley, P. D. Herbivory and defensive characteristics of tree species in a lowland tropical forest. *Ecol. Monogr.* **53**, 209–233 (1983).
- Korine, C., Kalko, E. K. V. & Herre, E. A. Fruit characteristics and factors affecting fruit removal in a Panamanian community of strangler figs. *Oecologia* **123**, 560–568 (2000).
- Fischer, K. E. & Chapman, C. A. Frugivores and fruit syndromes: differences in patterns at the genus and species level. *Oikos* **66**, 472–482 (1993).

19. Terborgh, J. *Five New World Primates* (Princeton Univ. Press, Princeton, 1983).
20. Leighton, M. Modeling dietary selectivity by Bornean orangutans: evidence for integration of multiple criteria in fruit selection. *Int. J. Primatol.* **14**, 257–313 (1993).
21. Gautier-Hion, A. *et al.* Fruit characters as a basis of fruit choice and seed dispersal in a tropical forest vertebrate community. *Oecologia* **65**, 324–337 (1985).
22. Davies, A. G. & Oates, J. F. in *Colobine Monkeys* (eds Davies, A. G. & Oates, J. F.) 229–249 (Cambridge Univ. Press, Cambridge, 1994).
23. Wrangham, R. W., Conklin-Brittain, N. L., & Hunt, K. D. Dietary responses of chimpanzees and cercopithecines to seasonal variation in fruit abundance: I. antifeedants. *Int. J. Primatol.* **19**, 949–970 (1998).
24. Onishi, A. *et al.* Dichromatism in macaque monkeys. *Nature* **402**, 139–140 (1999).
25. Struhsaker, T. T. *Ecology of an African Rainforest* (Univ. Florida Press, Gainesville, 1997).
26. Lucas, P. W. *et al.* Fieldkit to characterize the physical, chemical, and spatial aspects of potential primate foods. *Folia Primatol.* **72**, 11–25 (2001).
27. Osorio, D. & Vorobyev, M. Colour vision as an adaptation to frugivory in primates. *Proc. R. Soc. Lond. B* **263**, 593–599 (1996).
28. Darvell, B. W., Lee, P. K. D., Yuen, T. D. B., Lucas, P. W. *Meas. Sci. Technol.* **7**, 954–962 (1996).
29. Newton-Fisher, N. E. The diet of chimpanzees in the Budongo Forest Reserve, Uganda. *Afr. J. Ecol.* **37**, 344–354 (1999).
30. Gartlan, J. S., McKey, D. B., Waterman, P. G., Mbi, C. N. & Struhsaker, T. T. A comparative study of the phytochemistry of two African rainforests. *Biochem. Syst. Ecol.* **8**, 401–422 (1980).

Acknowledgements

We thank D. Osorio for help with colour registration; E. Ting, P. Y. Cheng, I. C. Bruce, R. T. Corlett, L. Ramsden, N. Yamashita and A. Walker for comments, P. Kagoro, B. Balyeganira and M. Musana for field assistance in Uganda; J. Magnay, R. W. Wrangham and C. A. Chapman for logistic support in Uganda; and the Ugandan National Council for Science and Technology, Ugandan Wildlife Authority and Makerere University Biological Field Station for permission to work at Kibale. Supported by Research Grants Council of Hong Kong, National Geographic Society, Sigma Xi, Explorer's Club and Croucher Foundation of Hong Kong.

Correspondence and requests for materials should be addressed to N.J.D. (e-mail: njdominy@hkusua.hku.hk).

Suppressing unwanted memories by executive control

Michael C. Anderson & Collin Green

Department of Psychology, University of Oregon, Eugene, Oregon 97403-1227, USA

Freud proposed that unwanted memories can be forgotten by pushing them into the unconscious, a process called repression¹. The existence of repression has remained controversial for more than a century, in part because of its strong coupling with trauma, and the ethical and practical difficulties of studying such processes in controlled experiments. However, behavioural and neurobiological research on memory and attention shows that people have executive control processes directed at minimizing perceptual distraction^{2,3}, overcoming interference during short and long-term memory tasks^{3–7} and stopping strong habitual responses to stimuli^{8–13}. Here we show that these mechanisms can be recruited to prevent unwanted declarative memories from entering awareness, and that this cognitive act has enduring consequences for the rejected memories. When people encounter cues that remind them of an unwanted memory and they consistently try to prevent awareness of it, the later recall of the rejected memory becomes more difficult. The forgetting increases with the number of times the memory is avoided, resists incentives for accurate recall and is caused by processes that suppress the memory itself. These results show that executive control processes not uniquely tied to trauma may provide a viable model for repression.

Executive control processes studied in behavioural^{6,9,14} and neurobiological^{2,4,10–13,15–17} research on cognition may provide a mechanism for the voluntary form of repression (suppression)

proposed by Freud¹. To test this hypothesis, we adapted the go/no-go paradigm used to study executive control over motor actions in primates¹⁸ and humans^{15–17} for use in a memory retrieval task. First, we trained subjects on 40 unrelated word pairs (for example, ordeal–roach) so that they could recall the right-hand member of each pair when provided with the left-hand member. Next, subjects performed a critical task requiring them to exert executive control over the retrieval process. On each trial of this think/no-think task, a cue from one of the pairs appeared on the computer screen. Depending on which cue appeared, subjects were told either to recall and say (think about) the associated response word (respond pairs), or not to think about the response (suppression pairs). For the latter pairs, we emphasized that subjects should not allow the associated memory to enter consciousness at all. If subjects accidentally responded to a suppression pair, they heard a beep signalling an error. To increase the need to recruit inhibitory control mechanisms, we required subjects to fixate on the cue word for the entire time (4 s) that it appeared on the screen, discouraging perceptual avoidance and generating a constant threat that the associated memory might intrude into consciousness. Thus, suppression trials required the stopping of both a prepotent motor

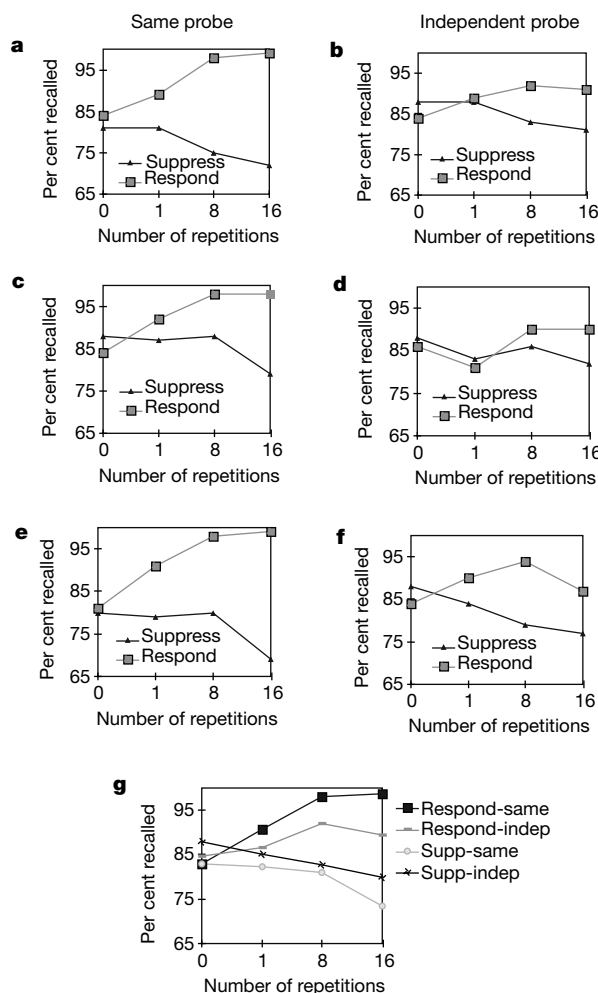


Figure 1 Final recall for respond and suppression items as a function of the number of repetitions for the same-probe (SP) and independent-probe (IP) tests. **a, b**, Experiment 1; **c, d**, experiment 2; **e, f**, experiment 3; **g**, averaged across experiments. Note the negative slope for recall of the suppressed item, indicating increasing inhibition. Inhibition (0 vs 16 suppressions) was significant ($P < 0.01$) in all experiments, and did not interact with type of test cue ($F < 1$ in all cases; analysis of variance). Inhibition was significant ($P < 0.05$) in every SP and IP test for every experiment (**a–f**).

response and the entrance of an unwanted memory into awareness. Before this phase, we told subjects which cues would require suppression so that they would recognize these words on sight. Subjects performed suppression and respond trials on different pairs that were intermixed. On each trial, no visual marking indicated which cue words were suppression items, forcing subjects to identify each cue to know whether to recall or suppress the associated memory.

The objective of the think/no-think task was to determine whether attempting to prevent awareness of an unwanted memory would hinder its later retrieval. To evaluate whether this occurred, the next phase required subjects to recall the response for each of the cue words. We emphasized that the previous goal of avoiding the associated items was no longer relevant and that a response should be recalled for every cue. If trying to prevent a memory from entering awareness recruits inhibitory control processes that impair that memory's retrievability, recall for suppression items should be worse than for baseline pairs on this test. In all experiments reported here, baseline items were studied pairs that did not appear as either respond or suppression items during the think/no-think phase.

In experiment 1, we varied the number of suppression or respond trials given for each pair. Subjects received 0, 1, 8 or 16 trials for each suppression and each respond item. If excluding unwanted memories from awareness recruits inhibitory control, recall should be worse after 16 suppression attempts than after 0 attempts (baseline). As shown in Fig. 1a, final recall of suppressed items was worse than of baseline items, and impairment increased linearly with suppression practice. In contrast, recall improved across repetitions for respond items, demonstrating the established benefits of retrieving memories on their later recall. These diverging patterns show that controlling awareness not only terminated the powerful facilitative effects of retrieval, but also impaired the recall of suppressed items to below their baseline level (0 suppression attempts). The increasing inhibition with repetition further indicates that unwanted memories might be especially vulnerable in settings requiring protracted avoidance, unlike the modest time (1 min over 16 suppressions) afforded by our task.

Impaired memory for suppression items indicates that there may be an executive control process that suppresses (reduces activation of) the unwanted memory itself (for example, roach in Fig. 2). However, mechanisms other than inhibition may be at work⁶. For instance, one strategy for avoiding an unwanted memory would be to generate diversionary thoughts to environmental stimuli that

remind us of it. New associations between these stimuli and the diversionary thoughts may interfere during later attempts to recall the memory. Alternatively, terminating retrieval may degrade the association between the cue and the response. Neither of these alternatives requires us to assume that the unwanted memory itself is inhibited and thus they do not require the postulation of an executive inhibition process. To isolate the contribution of inhibition, we used the independent probe method⁶. If inhibition impairs the unwanted memory itself, recall should be worse regardless of whether that item is tested with the cue used to train suppression or with a novel cue (Fig. 2). However, associative interference and unlearning predict that forgetting should be limited to the originally trained cue. To distinguish between these models, we retested subjects from experiment 1 with cues not previously encountered in the experiment ('independent probes'). For each item, we cued subjects with a semantic category and the initial letter of the response word (for example, for ordeal—roach, subjects received 'insect r__') and asked them to recall the studied word that fit the cues. On this new independent probe (IP) test, recall of suppressed items was again worse than baseline (Fig. 1b), and the impairment was higher when recall had been avoided more. The amount of forgetting did not differ reliably from when the originally trained cue was used in the test (for example, ordeal in the same probe (SP) test condition, Fig. 1a). This finding rules out associative interference and unlearning and shows that impairment is localized to the unwanted memory itself. This strongly supports the existence of an inhibitory control mechanism⁶.

It was necessary to show that subjects did not recall and then withhold suppression items on the final test. They might have done so out of confusion, given our emphasis on withholding responses during the think/no-think phase. To address this, we expanded our test instructions to emphasize that subjects should recall an item to every cue regardless of earlier instructions, even if they were guessing. To further encourage responding, we offered a monetary reward for correct answers. These incentives had little impact on the inhibition effect (Fig. 1c, d). Impairment was significant overall, increased with repetition, and did not vary with the type of test cue, again supporting the inhibitory control view. Although the instructions to guess enhanced recall of suppression and baseline items when the studied cue was given ($P < 0.01$), the linear trend for inhibition did not differ reliably across experiments ($F < 1$).

Subjects might have guessed the hypothesis of the experiment and tried to conform to the expectation of impaired memory by withholding items. To address this, we altered the final test instructions

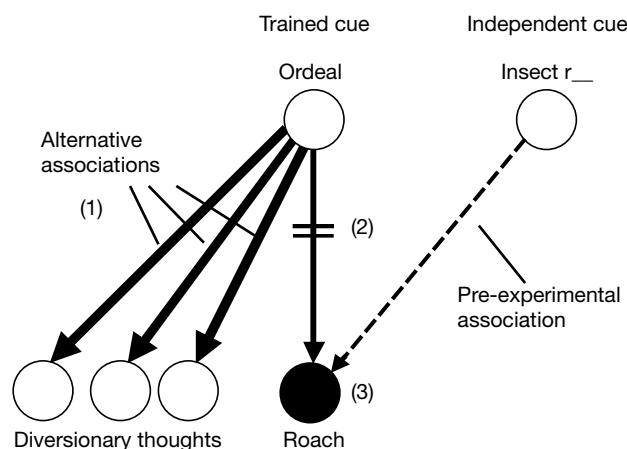


Figure 2 Three mechanisms that can explain impaired recall in the same-probe condition, illustrated with a stimulus pair. Associative interference posits that suppression training leads subjects to generate diversionary thoughts (1) to the trained cue that interfere during later attempts to recall the target. Unlearning assumes that suppression training weakens

the cue–target connection (2). The suppression hypothesis states that suppression training inhibits the target (3). Note that testing the target with an independent cue circumvents interference (1) and unlearning (2). Any impairment found with this test may be attributed to effects localized to the target.

to make subjects believe that we expected improved memory for suppression items. We told subjects that research suggests that when people try not to think about something, they ironically think about it more, as when people try not to think about falling asleep at night and then experience insomnia¹⁹. We noted that this research predicts that memory for suppression items should improve because the avoided memories should intrude during suppression trials. Post-experimental questionnaires indicated that subjects believed that this was our rationale (average rating of 4.2 on a 5-point belief scale) and most endorsed the theory on the basis of experience. These new test instructions had little effect on the inhibition pattern (Fig. 1e, f).

Success in our think/no-think task is defined by participants' subjective awareness of the unwanted memory during suppression trials, a state which cannot be measured. Efforts to control this state nevertheless left their mark on memory. However, the apparent memory deficit may reflect suppression of the overt motor response associated with the unwanted memory. To test this, we removed the instruction to suppress awareness of the memory during the think/no-think phase. We instead asked subjects to recall the memory, but not say it aloud, rendering this an episodic go/no-go task. Recall did not decline with the number of verbal suppressions, and this pattern did not vary with the cue used to test items (Fig. 3a, b). In contrast, when experiments 1–3 are combined, the linear trend for inhibition was reliable for the SP and IP test conditions ($P < 0.001$, Fig. 1g) and these effects did not vary reliably across experiments. Importantly, the linear inhibition effect was statistically larger in the think/no-think task (experiments 1–3) than in the go/no-go task (experiment 4). These findings isolate impaired recall of unwanted memories to cognitive operations directed at keeping them out of awareness.

Our results imply that a process exists that impairs the retention of memories when they are deliberately kept out of consciousness. When people encounter a stimulus that is known to cue an unwanted memory, this process can be recruited to prevent awareness of the memory. The regulation of consciousness is accomplished by an inhibitory control mechanism that suppresses the unwanted memory itself (as shown by our independent probe data), and not merely by the momentary filling of working memory with diversionary thoughts. Delayed recall of the unwanted memory is worse than a baseline condition in which no reminders of the memory were presented in the interim. This paradoxical pattern arises because repeated exposure to reminders makes it necessary for people to adapt their patterns of thought internally with executive control processes. Thus, frequent encounters with reminders should make an unwanted memory less accessible, a finding

noted in some clinical studies of psychogenic amnesia²⁰.

Our results clarify the nature and scope of inhibitory processes previously shown to impair episodic recall, and link those processes to behavioural and neurobiological research on executive control. Work on retrieval-induced forgetting⁶ has shown that memories that interfere with the retrieval of other targets are inhibited, but it has not previously been established whether inhibition is under strategic control. Research into directed forgetting^{7,21,22} suggests a controllable inhibition process, but one limited in scope to an immediately preceding temporal interval. The present findings show a controllable inhibition process that can be flexibly targeted to a specific prepotent memory after intervening memories have been acquired. These findings do not support the popular idea that attempting to suppress an unwanted thought makes it hyperaccessible, at least as this idea applies to suppressing episodic memories¹⁹.

The need to terminate retrieval in the think/no-think paradigm can be viewed as one instance of the need to override prepotent or habitual responses when they are inappropriate, a function presumed to require executive control^{4,8–10,12,13,23,24}. Related tasks such as the go/no-go procedure are widely used to study this executive function and have been shown to recruit dorsolateral prefrontal cortex (DLPFC) and anterior cingulate cortex (ACC) (with ventral PFC)^{15–17}. Other studies show a specific role of DLPFC in overcoming interference from competing representations in working memory⁴, in selecting one item in working memory as the basis for responding²⁶ and in on-line manipulation of information²⁷. The sustained regulation of awareness required by our procedure is likely to use such processes. This suggests that the think/no-think task recruited DLPFC and ACC, perhaps with medial temporal regions, to control awareness of the memory^{28,29}. If this is true, the present memory deficits provide a behavioural marker of the action of this network on rejected memories and suggest that this network exerts control through inhibition^{4,6,11,12,25}.

Whatever their neural basis, our results establish a direct link between internal operations that control phenomenal awareness of a memory and its later accessibility. These findings thus support a suppression mechanism that pushes unwanted memories out of awareness, as posited by Freud. Research on retrieval-induced forgetting shows that suppression can have a lasting impact on a memory's accessibility⁶. Suppression in the current paradigm may be similarly enduring. Furthermore, if retrieving diversionary thoughts becomes habitual, inhibition may be sustained without any intention of avoiding the unwanted memory³⁰. Together, these factors provide a viable model of repression, and its potential evolution from an intentional to an unintentional process. If so, repression may not be tied uniquely to psychological defence, but may rather reflect the action of a general executive control process, directed at declarative memories of past experience. □

Methods

Thirty-two neurologically normal college students participated in each experiment, except experiment 4, in which there were sixteen. Each participant was trained on 40 critical and 10 filler word pairs. The stimulus and response members of each pair had a weak pre-experimental relationship, and were unrelated to words in other pairs. The response members were chosen so that each was a member of its own category (for example, insects), to permit later testing of that item with an extralist category cue. The word pairs were exposed individually for 5 s in the centre of a computer screen with the response printed to the right of the stimulus. Test–feedback cycles followed in which subjects were presented with the stimulus member and asked to say the response aloud as quickly as possible. The correct answer was given visually if the response was omitted. Test–feedback cycles on all the pairs continued until a minimum of 50% of the pairs were correctly recalled.

After initial study, subjects were given the think/no-think phase instructions, and were then presented with the 15 stimulus members from the to-be-suppressed pairs without their responses. Subjects familiarized themselves with the stimuli so that they could identify them during suppression trials and prevent the associated memory item from coming to mind. After a brief practice session on the think/no-think task using filler items, subjects were given 377 trials in which respond and suppression stimuli were randomly

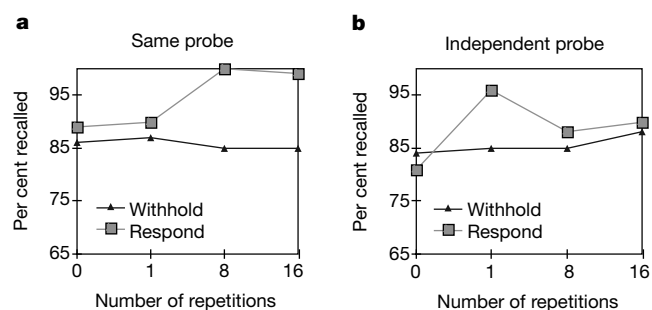


Figure 3 Final recall of withhold and respond items as a function of the number of repetitions. **a**, Performance in the same test probe (SP) condition. **b**, Performance in the independent probe condition (IP). In both conditions, withhold item performance does not decrease with repetition. None of the inhibition effects were significant ($F < 1$ in all cases). There was reliably more inhibition in experiments 1–3 than in experiment 4 ($P < 0.05$), and a reliably greater linear trend for inhibition ($P < 0.05$). Neither of these tendencies interacted with type of test cue ($F < 1$).

intermixed. On each trial, a fixation cross appeared for 200 ms, followed by a stimulus member in the centre of the screen. For respond trials, the stimulus was presented for up to 4 s, or until the subject responded, and subjects had to report the response as quickly as possible. For suppression trials, the stimulus remained on the screen for 4 s. If a subject responded, a loud error beep sounded. Trials were separated by a 400-ms intertrial interval. Suppression and respond trials were conducted on different word pairs, with five pairs participating in each of the 0 (baseline), 1, 8 and 16-repetition conditions for both the respond and suppression conditions. Respond trials on filler pairs were also included so that 67% of the trials in the think/no-think phase required a response, encouraging a strong mental set to respond that had to be overridden, as in go/no-go tasks.

After the think/no-think phase, we tested the subjects' memory for all of the word pairs in two ways. In both tests, a cue for each word pair was presented in the centre of the screen for up to 4 s, or until subjects spoke the response. Pairs for the different respond and suppress conditions were intermixed pseudo-randomly, with the constraint that the average test position for each condition was equated. In the same-probe test, subjects' recall for each pair was cued with the stimulus member that was paired with the response throughout the experiment. In the independent probe test, subjects were cued with the category name for each exemplar along with its first letter. In each case, subjects were asked to recall the studied item that fit the cues and not to withhold any items. Half of the subjects in each experiment got the same-probe test first, and half were given the independent probe test first.

Subjects in experiment 2 were given 25 cents for each correct answer, up to a maximum of \$4. Subjects in experiment 3 were given a questionnaire after the experiment in which they were asked about their impressions of our (false) hypothesis that memory would improve with attempts to suppress an item.

Received 10 October; accepted 21 December 2000.

- Freud, S. in *The Standard Edition of the Complete Psychological Works of Sigmund Freud 1* (ed. J. Strachey) 117–128 (Hogarth, London, 1966).
- Chao, L. L. & Knight, R. T. Human prefrontal lesions increase distractibility to irrelevant sensory inputs. *Cog. Neurosci. Neuropsychol.* **6**, 1605–1610 (1995).
- Dagenbach, D. & Carr, T. H. (eds) *Inhibitory Processes in Attention, Memory, and Language* (Academic, San Diego, 1994).
- Smith, E. E. & Jonides, J. Storage and executive processes in the frontal lobes. *Science* **283**, 1657–1661 (1999).
- Hasher, L. & Zacks, R. T. Working memory, comprehension and aging: A review and a new view. *Psychol. Learn. Motiv.* **22**, 193–225 (1988).
- Anderson, M. C. & Spellman, B. A. On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychol. Rev.* **102**, 68–100 (1995).
- Bjork, R. A. in *Varieties of Memory and Consciousness: Essays in Honour of Endel Tulving* (eds Roediger, H. L. & Craik, F. I. M.) 309–330 (Lawrence Erlbaum Associates, Hillsdale, 1989).
- Luria, A. R. *Higher Cortical Function in Man* (Basic Books, New York, 1966).
- Logan, G. D. & Cowan, W. B. On the ability to inhibit thought and action: A theory of an act of control. *Psychol. Rev.* **91**, 295–327 (1984).
- Posner, M. I. & Peterson, S. E. The attention system of the human brain. *Annu. Rev. Neurosci.* **13**, 25–42 (1990).
- Knight, R. T., Staines, W. R., Swick, D. & Chao, L. L. Prefrontal cortex regulates inhibition and excitation in distributed neural networks. *Acta Psychol.* **101**, 159–178 (1999).
- Cohen, J. D. & Servan-Schreiber, D. Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychol. Rev.* **99**, 45–77 (1992).
- Carter, C. S., Botvinick, M. M. & Cohen, J. D. The contribution of the anterior cingulate cortex to executive processes in cognition. *Rev. Neurosci.* **10**, 49–57 (1999).
- Mayr, U. & Keele, S. W. Changing internal constraints on action: The role of backward inhibition. *J. Exp. Psychol. Gen.* **129**, 4–26 (2000).
- Casey, B. J. et al. A developmental functional MRI study of prefrontal activation during performance of a go-no-go task. *J. Cogn. Neurosci.* **9**, 835–847 (1997).
- Garavan, H., Ross, T. J. & Stein, E. A. Right hemispheric dominance of inhibitory control: An event-related functional MRI study. *Proc. Natl Acad. Sci. USA* **96**, 8301–8306 (1999).
- de Zubicaray, G. I. et al. Motor response suppression and the prepotent tendency to respond: A parametric fMRI study. *Neuropsychologia* **38**, 1280–1291 (2000).
- Sakagami, M. & Niki, H. Spatial selectivity of go/no go neurons in the monkey prefrontal cortex. *Exp. Brain Res.* **100**, 165–169 (1994).
- Wegner, D. M. Ironic processes of mental control. *Psychol. Rev.* **101**, 34–52 (1994).
- Freyd, J. J. *Betrayal Trauma: The Logic of Forgetting Childhood Abuse* (Harvard Univ. Press, Cambridge, MA, 1996).
- Geiselman, R. E., Bjork, R. A. & Fishman, E. L. Disrupted retrieval in directed forgetting: A link with posthypnotic amnesia. *J. Exp. Psychol. Gen.* **112**, 58–72 (1983).
- Conway, M. A., Harries, K., Noyes, J., Racsmany, M. & Frankish, C. R. The disruption and dissolution of directed forgetting: Inhibitory control of memory. *J. Mem. Lang.* **43**, 409–430 (2000).
- Norman, D. A. & Shallice, T. in *Consciousness and Self-Regulation: Advances in Research and Theory* (eds Davison, R. J., Schwartz, G. E. & Shapiro, D.) 1–18 (Plenum, New York, 1986).
- MacDonald, A. W., Cohen, J. D., Andrew-Stenger, V. & Carter, C. S. Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science* **288**, 1835–1838 (2000).
- Shimamura, A. P. The role of the prefrontal cortex in dynamic filtering. *Psychobiology* **28**, 207–218 (2000).
- Rowe, J. B. et al. The prefrontal cortex: Response selection or maintenance within working memory. *Science* **288**, 1656–1660 (2000).
- D'Esposito, M. et al. Functional MRI studies of spatial and nonspatial working memory. *Cogn. Brain Res.* **7**, 1–13 (1998).
- Schacter, D. L. & Wagner, A. D. Medial temporal lobe activations in fMRI and PET studies of episodic encoding and retrieval. *Hippocampus* **9**, 7–24 (1999).
- Schacter, D. L. Memory and awareness. *Science* **280**, 59–60 (1998).

- Anderson, M. C. Active Forgetting: Evidence for functional inhibition as a source of memory failure. *J. Aggression Maltreatment Trauma* (in the press).

Acknowledgements

The research reported here was supported by a grant from the US National Institute of Mental Health.

Correspondence and requests for materials should be addressed to M.C.A. (e-mail: mcanders@darkwing.uoregon.edu).

Masking unveils pre-amodal completion representation in visual search

Robert Rauschenberger & Steven Yantis

Department of Psychology, The Johns Hopkins University, Baltimore, Maryland 21218, USA

When one object is partly occluded by another, its occluded parts are perceptually 'filled in', that is, the occluded object appears to continue behind its occluder. This process is known as amodal completion¹. The completion of a partially occluded object takes about 200 ms (ref. 2), and pre-completion information (that is, information from before amodal completion has occurred) exists in the visual system for that duration^{2,3}. It has been suggested, however, that observers cannot make use of this information, even when it is beneficial to do so: visual search for a target that appears to be partly occluded, for example, is slower than for a target that does not undergo occlusion, despite both targets being physically identical^{4–6}. Here we show that visual search does have access to pre-completion representations, but only for a limited time that depends on the size of the occluded region.

Early investigations of visual search focused on discovering the elementary features available in early vision^{7–12}, whereas more recent work has demonstrated that the input to visual search is much more complex than previously assumed^{4–6,13,14}. Although it seems that the entry level for vision (that is, entire objects or individual features) can be quite high in many cases, questions remain about the nature of the information available at earlier stages of processing. In the case of amodal completion, for example, it seems that visual search relies on a post-completion representation even when this impairs search^{4–6}. This finding can be interpreted in at least three ways. First, it could be that there is no pre-amodal completion stage in processing. This is unlikely because other studies have shown that pre-completion information is available for certain perceptual judgements^{2,3}. Second, it is possible that pre-completion information exists only implicitly, as an 'ingredient' in the computation of a completed representation, and is not explicitly available for all perceptual judgements. Examples of this include monocular information during binocular rivalry¹⁵, and very high spatial frequency information¹⁶, both of which are present in

Table 1 Percentage of mean target-absent error rates

Experiment	Number of items in display			
	2	4	6	8
Medium notch (100-ms SOA)	11.4	17.7	20.3	12.3
Medium notch (250-ms SOA)	3.5	4.6	5.5	3.0
Small notch	18.7	19.8	22.5	15.6
Large notch	2.0	4.0	3.0	1.9

early vision but are apparently barred from awareness. The third possibility is that an explicit pre-completion representation is at least momentarily available, but a later amodally completed representation pre-empts the pre-completion representation in visual search¹⁴.

The third interpretation predicts that it should be possible to reveal the presence of the pre-completion representation by interrupting processing before amodal completion has been fully achieved. Such an interruption can be achieved by the use of a mask^{17,18}. In our experiments, participants searched for a notched disk target among complete disks and squares (Fig. 1). With unlimited exposure duration, when the notched target disk abuts a square (adjacent condition), search is inefficient because the notched target is rendered similar to the complete distractor disks by amodal completion; however, search is efficient when the target stands separate from the square (separate condition) and completion is not possible⁵. This result suggests that visual search mechanisms do not have access to the pre-completion representation. The same result should be obtained if a mask interrupts processing of the display after amodal completion has had time to occur (> 200 ms (ref. 2)). If, however, a mask terminates the display before amodal completion finishes (< 200 ms), search should be equally efficient in the separate and adjacent conditions because the representation of the target in the adjacent condition remains notched, and therefore is easily discriminated from the distractor disks. We used two different stimulus-onset asynchronies (SOAs) for the mask: 100 ms (intended to prevent amodal completion) and 250 ms (long enough to permit completion). All conditions were presented in mixed blocks to prevent the use of a stimulus- or SOA-dependent strategy.

The results are shown in Fig. 2; target-absent data for this and all subsequent experiments are summarized in Table 1—our analyses will centre on the target-present trials because the target-absent trials do not contain any candidates for amodal completion. Our statistical analyses focus on error rate. The interpretation of response times is problematic when error rates vary systematically across conditions. Search efficiency (as indexed by the slope of the search function) was greater in the separate condition than in the

adjacent condition for the 250-ms SOA (Fig. 2b; analysis of variance (ANOVA), $F(3, 33) = 4.84$, $P < 0.01$), which replicates previous findings⁵. In the 100 ms condition, there was no difference in search efficiency (Fig. 2a; $F(3, 33) = 1.03$, not significant). The absence of a difference in search efficiency in the 100 ms condition was not due to a decrease in search efficiency for the more efficient separate condition ($F < 1$). This result is paradoxical: a manipulation that increased task difficulty (that is, reducing the target-mask SOA) yielded increased search efficiency.

Although there was no difference in search efficiency in the 100-ms SOA condition, there was a main effect of target type (adjacent versus separate) ($F(1, 11) = 8.72$, $P < 0.05$). We attribute this effect, which has been reported elsewhere⁵, to differences between the two target types that are unrelated to amodal completion. Amodal completion affects the slopes of the search functions because it renders the target more similar to the distractors, incurring an additional cost with every additional distractor item in the display. The greater intercept in the adjacent condition, on the other hand, indicates factors affecting the target irrespective of the number of distractors. Such factors may include the greater spatial proximity of the target disk to its neighbouring square in the adjacent condition than in the separate condition, and hence its greater vulnerability to

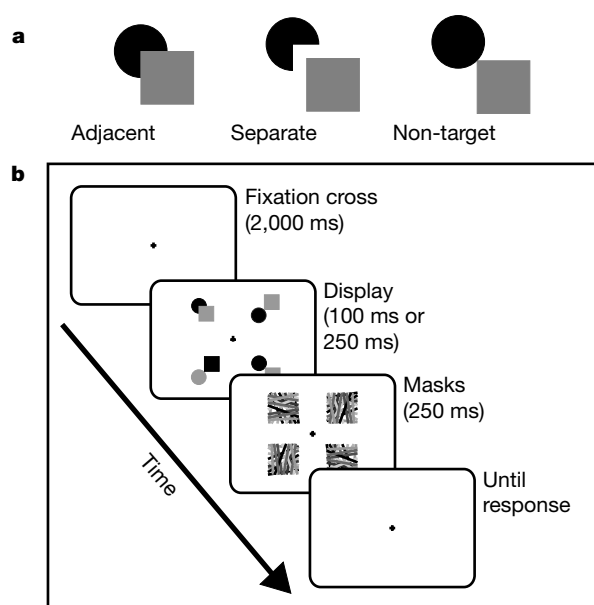


Figure 1 Visual search task. **a**, The target is either a notched disk abutting a square (adjacent condition) or standing separate from the square (separate condition); distractors are complete disks and squares. **b**, Sequence of events on each trial (see Methods for details).

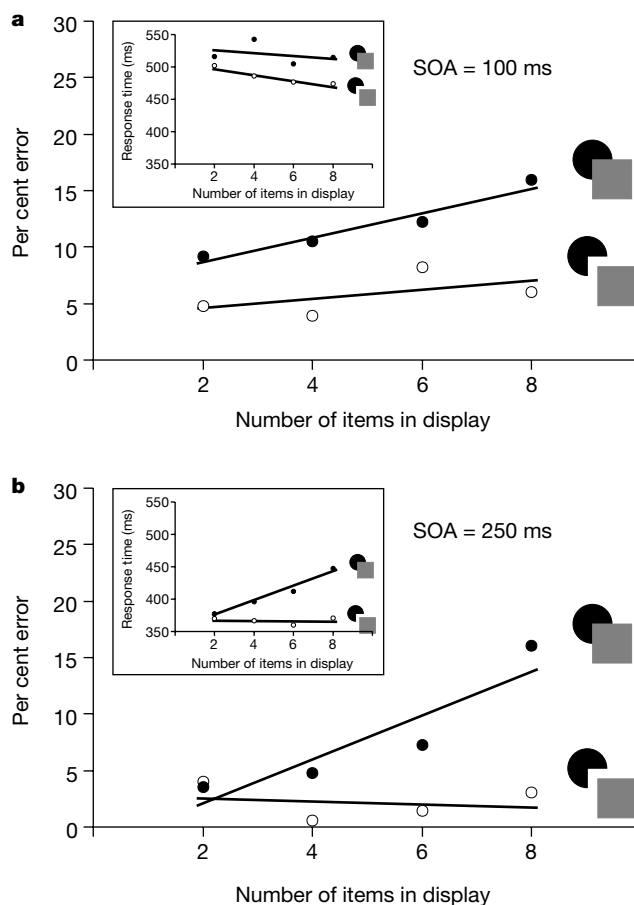


Figure 2 Participants' performance as a function of SOA, target type (adjacent or separate), and number of items in the display. Closed symbols represent the adjacent condition, open symbols the separate condition (also for Figs 3 and 4). The insets show the response-time data for each condition. **a**, 100-ms SOA. The adjacent and separate conditions yielded equivalent search efficiencies. **b**, 250-ms SOA. Search was less efficient in the adjacent than in the separate condition. The response-time data show the same pattern of results, with equally efficient search in the 100 ms condition and a pronounced slope difference in the 250 ms condition.

lateral masking or related effects.

Our use of error rates as a measure of search efficiency is validated by the congruence of the response-time data with the error data (see Fig. 2, insets), and by the fact that the 250-ms condition error data replicate previous findings⁵. As a further check on the reliability and validity of our results, we replicated the new finding from the first experiment (the equality of search efficiency for adjacent and separate targets at the 100-ms SOA) using response time as a dependent measure. The design of the experiment was the same as in the previous experiment, with the exception that the target and mask displays were cycled continuously until a response was issued, and that participants made a speeded response with accuracy being stressed. We assumed that the mask would prevent completion on each cycle. We recorded response time as a

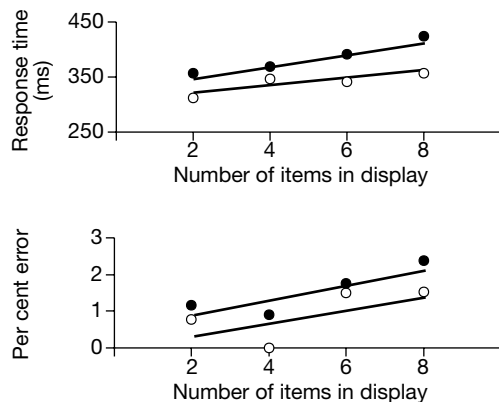


Figure 3 Participants' performance as a function of target type (adjacent or separate) and number of items in the display. A replication of the 100 ms condition using response time as a dependent measure confirmed the results of the first experiment.

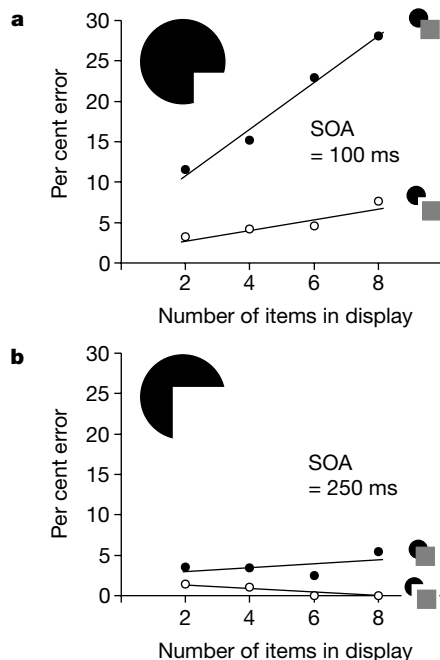


Figure 4 Participants' performance as a function of target type (adjacent or separate) and number of items in the display. **a**, Small-notched target disk. When a target disk with a small notch is used, there is a difference in search efficiency between the adjacent and separate conditions even at the 100-ms SOA. **b**, Large-notched target disk. When a target disk with a large notch is used, both target types yield equivalent search efficiencies even at the 250-ms SOA.

function of display exposure time, and excluding mask presentation time. The results are shown in Fig. 3. As before, the slopes for the adjacent and separate targets are comparable (7.5 versus 11.1 ms per item) and statistically indistinguishable ($t(14) = 1.53$, not significant; Student's t -test). This finding stands in marked contrast to the pattern of results observed with much longer display durations⁴⁻⁶.

We extended the result obtained in the first experiment by manipulating the size of the occluded region, which is known to affect the time required for amodal completion, with large regions requiring more time than smaller regions¹⁹. If the size of the notch in the target is reduced substantially (from 25 to 9% of the entire disk), and amodal completion can thus be achieved even before the onset of the mask at the 100-ms SOA¹⁹, we should observe a difference in search efficiency between the adjacent and separate conditions, even at this short SOA. If, conversely, the size of the notch is increased substantially (from 25 to 37%), amodal completion should be slowed to such an extent that even the 250-ms SOA will interrupt processing before the pre-completion representation is pre-empted. In this case, one would predict no difference in search efficiency at the long SOA.

As shown in Fig. 4, these predictions were borne out. The target with the small notch yielded a difference in search efficiency even at the 100-ms SOA ($F(3, 33) = 2.90$, $P < 0.05$). The target with the large notch, by contrast, produced no difference in search efficiency even at the 250-ms mask SOA ($F(2, 22) = 1.42$, not significant). The results for the target with the small notch (Fig. 4a) were qualitatively similar to those obtained in the first experiment in the 250 ms condition, as one would expect if amodal completion had occurred too quickly even for the 100-ms SOA. An ANOVA test comparing the two conditions across experiments confirmed that the results were statistically indistinguishable ($F < 1$). Similarly, the large-notch target yielded a pattern of results similar to that observed in the 100 ms condition of the first experiment (see Fig. 4b). Again, statistical analysis confirmed that the results were comparable ($F < 1$).

Our findings indicate not only that pre-amodal completion information is present in the visual system along the path to a final, completed representation, but that, at least briefly, its representation is sufficiently explicit to subserve overt, conscious behaviour. The assertion that visual search has no access to low-level information^{4-6,20} should be tempered by the finding that it can, under some conditions, gain access to this information. Of course, obligatory use of a post-completion representation under normal circumstances makes adaptive sense because search would otherwise be routinely inefficient in a world that abounds with occlusions. □

Methods

Participants performed visual search for a target that was either a notched disk abutting a square (adjacent condition), or a notched disk standing separate from the square (separate condition); distractors were complete disks and squares (see Fig. 1a). In the first case, the target was a candidate for amodal completion; in the second case, it was not. Displays consisted of 2, 4, 6 or 8 disk-square combination pairs arranged around a central fixation cross; on half the trials one of these was a target pair. The diameter of the disk subtended 1.2° of visual angle, and one side of the square subtended the same angle. The entire display measured 8.0° across.

Each trial began with a fixation cross at the centre of the screen, followed by the search display (see Fig. 1b). The search display was replaced by a mask display after either 100 or 250 ms; the duration of the mask display was 250 ms. The mask display was followed by a blank screen with a fixation cross. Participants reported whether the notched disk target was present or absent by pressing one of two keys. We stressed the need for accuracy over speed. Participants completed one practice block with a long mask-onset time (950 ms) and another with the actual parameters used in the subsequent experimental blocks (100 and 250 ms). All conditions were presented in fully mixed blocks. Participants completed a total of 512 experimental trials. A new set of 12 undergraduate students naive to the purposes of the study participated in each experiment. Fifteen observers participated in the response time replication of the new finding of the first experiment. In this experiment, the target and mask displays were cycled repeatedly, without inter-stimulus interval, until a response was made.

Received 17 October; accepted 22 December 2000.

1. Michotte, A., Thines, G. & Crabbé, G. in *Michotte's Experimental Phenomenology of Perception* (eds G. Thines, A. Costall & G. Butterworth) (Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1991).
2. Sekuler, A. B. & Palmer, S. E. Perception of partly occluded objects: a microgenetic analysis. *J. Exp. Psychol. Gen.* **121**, 95–111 (1992).
3. Gerbino, W. & Salmaso, D. The effect of amodal completion on visual matching. *Acta Psychol.* **65**, 25–46 (1987).
4. He, Z. J. & Nakayama, K. Surfaces versus features in visual search. *Nature* **359**, 231–233 (1992).
5. Rensink, R. A. & Enns, J. T. Early completion of occluded objects. *Vis. Res.* **28**, 2489–2505 (1998).
6. Davis, G. & Driver, J. Kanizsa subjective figures can act as occluding surfaces at parallel stages of visual search. *J. Exp. Psychol. Hum. Percept. Perform.* **24**, 169–184 (1998).
7. Treisman, A. & Gelade, G. A feature integration theory of attention. *Cogn. Psychol.* **12**, 97–136 (1980).
8. Treisman, A. Preattentive processing in vision. *Comp. Vis. Graphics Image Proc.* **31**, 156–177 (1985).
9. Wolfe, J. M., Cave, K. R. & Franzel, S. L. Guided search: an alternative to the feature integration model for visual search. *J. Exp. Psychol. Hum. Percept. Perform.* **15**, 419–433 (1989).
10. Wolfe, J. M. Guided search 2.0: a revised model of visual search. *Psychonom. Bull. Rev.* **1**, 202–238 (1994).
11. Wolfe, J. M. in *Attention* (ed. Pashler, H.) 13–73 (Psychology Press, Hove, 1998).
12. Julesz, B. Textons, the elements of texture-perception, and their interactions. *Nature* **290**, 91–97 (1981).
13. Donnelly, N., Humphreys, G. W. & Riddoch, M. J. Parallel computation of shape description. *J. Exp. Psychol. Hum. Percept. Perform.* **17**, 561–570 (1991).
14. Rensink, R. A. & Enns, J. T. Preattentive effects in visual search: Evidence for low-level grouping. *Psychol. Rev.* **102**, 101–130 (1995).
15. Blake, R. What can be 'perceived' in the absence of visual awareness? *Curr. Dir. Psychol. Sci.* **6**, 157–162 (1997).
16. Crick, F. & Koch, C. Are we aware of neural activity in primary visual cortex? *Nature* **375**, 121–123 (1995).
17. Breitmeyer, B. G. *Visual Masking: An Integrative Approach* (Oxford Univ. Press, New York, 1984).
18. Enns, J. T. & Di Lollo, V. What's new in visual masking? *Trends Cogn. Sci.* **4**, 345–352 (2000).
19. Shore, D. I. & Enns, J. T. Shape completion time depends on the size of the occluded region. *J. Exp. Psychol. Hum. Percept. Perform.* **23**, 980–998 (1997).
20. Wolfe, J. M. & Horowitz, T. S. A new look at preattentive vision. *Invest. Ophthalmol. Vis. Sci.* **39**, S872 (1998).

Acknowledgements

This work was supported by a grant from the National Institutes of Health. We are grateful to J. Enns and J. Wolfe (who suggested the response-time control experiment) for their comments on an earlier draft.

Correspondence and requests for materials should be addressed to R.R. (e-mail: rrausch@jhu.edu) or S.Y. (e-mail: yantis@jhu.edu).

Neurogenesis in the adult is involved in the formation of trace memories

Tracey J. Shors*, George Miesegaes*, Anna Beylin†, Mingrui Zhao*, Tracy Rydel† & Elizabeth Gould†

* Department of Psychology and Center for Collaborative Neuroscience, Rutgers University, Piscataway, New Jersey 08854, USA

† Department of Psychology, Princeton University, Princeton, New Jersey 08544, USA

The vertebrate brain continues to produce new neurons throughout life^{1–12}. In the rat hippocampus, several thousand are produced each day, many of which die within weeks¹³. Associative learning can enhance their survival^{13,14}; however, until now it was unknown whether new neurons are involved in memory formation. Here we show that a substantial reduction in the number of newly generated neurons in the adult rat impairs hippocampal-dependent trace conditioning, a task in which an animal must associate stimuli that are separated in time¹⁵. A similar reduction did not affect learning when the same stimuli are not separated in time, a task that is hippocampal-independent^{16,17}. The reduction in neurogenesis did not induce death of mature hippocampal neurons or permanently alter neurophysiological properties of

the CA1 region, such as long-term potentiation. Moreover, recovery of cell production was associated with the ability to acquire trace memories. These results indicate that newly generated neurons in the adult are not only affected by the formation of a hippocampal-dependent memory¹³, but also participate in it.

We used a toxin for proliferating cells, the DNA methylating agent methylazoxymethanol acetate (MAM)^{18–20}, to diminish the number of adult-generated cells in the dentate gyrus of rats. First, we identified a dose that would deplete the new cells without impairing overall health. Adult male Sprague–Dawley rats ($n = 18$, 3 per dose) were injected subcutaneously with MAM at daily doses of 0, 3, 4, 5, 7 and 15 mg kg⁻¹ for 14 d. We chose this regime because numbers of newly generated cells increase up to 10 d after mitosis and decline thereafter²¹. Thus, 14 d of treatment should prevent their incorporation to the granule cell layer. Twelve hours after the last MAM injection, rats were injected with 5-bromodeoxyuridine (BrdU), a marker of dividing cells²², and perfused 2 h later (see Supplementary Information). Daily treatment for 14 d with 5 mg kg⁻¹ MAM decreased the number of granule cells labelled with BrdU and TUC-4, a marker of immature neurons²³, by roughly 80% ($P < 0.05$). Rats treated with 5 mg kg⁻¹ MAM continued to gain weight during treatment and were indistinguishable in general appearance from vehicle-treated rats. Liver cytology was normal. Doses lower than 5 mg kg⁻¹ had smaller effects on cell number, 7 mg kg⁻¹ was associated with weight loss, and 15 mg kg⁻¹ with weight loss and decline in overall health. Therefore, in the subsequent studies, rats were treated with 5 mg per kg per day, for 14 d.

In the first conditioning experiment, rats were fitted with headstages and electrodes to deliver periorbital stimulation to the eyelid and to record electromyographic (EMG) activity associated with eyelid closure. After 5 days of recovery, rats were injected subcutaneously with 5 mg kg⁻¹ MAM or saline daily, for 14 d. Rats received three injections of BrdU, on days 10, 12 and 14 (10, 8 and 6 d before perfusion) to label large numbers of immature cells. Rats were untreated for 2 d, then trained with delay or trace conditioning. In both tasks, a white noise conditioned stimulus predicts the occurrence of an unconditioned stimulus of periorbital eyelid stimulation. As the animal learns, it blinks in response to the conditioned

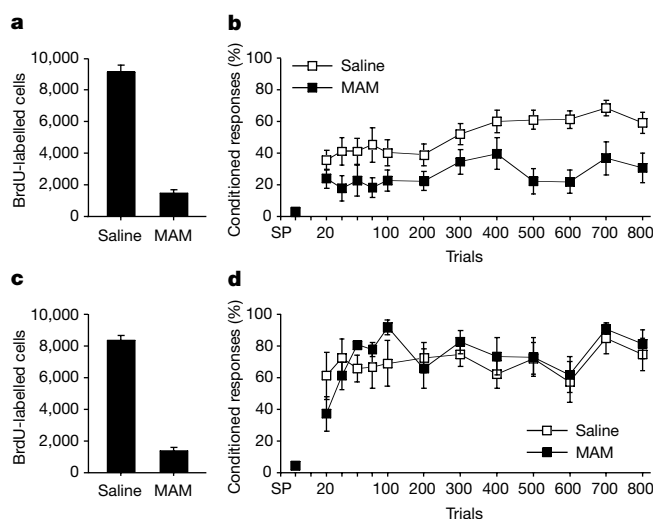


Figure 1 Adult-generated neurons are involved in trace but not in delay conditioning. **a**, Total numbers of BrdU-labelled cells in the dentate gyrus before trace conditioning. Bars represent mean $\pm$ s.e.m. **b**, Spontaneous blink rate (SP) and the percentage of conditioned responses (mean $\pm$ s.e.m.) across 800 trials of hippocampal-dependent trace conditioning in rats treated with MAM versus saline. **c**, Total numbers of BrdU-labelled cells in the dentate gyrus before delay conditioning. **d**, SP and percentage of conditioned responses across 800 trials of hippocampal-independent delay conditioning in rats treated with MAM versus saline.

stimulus. During delay conditioning, the conditioned stimulus and unconditioned stimulus overlap, and acquisition does not require an intact hippocampus^{16,17}. During trace conditioning, there is a temporal gap between the conditioned stimulus and the unconditioned stimulus, and the animal maintains or resurrects a memory 'trace' of the conditioned stimulus to associate it with the unconditioned stimulus. Acquisition of trace conditioning requires the hippocampus^{15,24,25}. The interstimulus interval was 750 ms for both delay and trace conditioning, and all rats were trained with 800 trials of paired stimuli, 200 trials a day for 4 d.

Treatment with MAM for 14 d reduced the number of newly generated cells in the dentate gyrus by 84% (1,487 for MAM ($n = 10$) and 9,140 for saline ($n = 12$)) ($P < 0.00001$) (Figs 1a and 2a, c). Most BrdU-labelled cells co-expressed markers TuJ1 (Fig. 2b), a marker for mature and immature neurons, and NeuN (Fig. 2d), a marker for mature neurons. MAM treatment reduced the overall number of conditioned responses during trace conditioning ($F(1,20) = 7.62$, $P = 0.01$) (Fig. 2b). Treatment did not alter spontaneous blink rate ($P > 0.05$), responses to the conditioned stimulus before training ($P > 0.05$), or responses on the first trial

($P > 0.05$). Treatment of a second group with MAM reduced the number of newly generated cells by 84% (1,357 for MAM ($n = 5$) and 8,334 for saline ($n = 6$)) ($P < 0.00001$) (Fig. 1c), yet MAM-treated rats rapidly acquired the hippocampal-independent task of delay conditioning. Treatment with MAM did not alter the number of conditioned responses ($F(1,9) = 0.15$, $P > 0.05$) or conditioned responses across trials ($F(12,99) = 1.28$, $P > 0.05$) (Fig. 1d). Treatment with MAM did not affect spontaneous blink rate ($P > 0.05$), responses to the conditioned stimulus before ($P > 0.05$) or on the first trial of training ($P > 0.05$).

In a second conditioning experiment, we replicated the detrimental effect of MAM treatment on trace conditioning and evaluated alterations in neuronal responsiveness and long-term plasticity in the hippocampus. Groups were injected with MAM (5 mg kg^{-1} , $n = 6$) or saline ($n = 6$) for 14 d, and trained 2 d later with 800 trials of trace conditioning. MAM reduced the number of newly generated cells by 75% (2,284 for MAM and 9,264 for saline) and reduced the number of conditioned responses during trace conditioning ($F(1,10) = 8.93$, $P = 0.01$). Rats treated with saline emitted $53 \pm 8\%$ conditioned responses over 800 trials whereas those treated with MAM emitted $19 \pm 9\%$ conditioned responses. *In vivo* recordings of extracellular excitatory postsynaptic potentials (EPSPs) were conducted in area CA1 of the hippocampus in response to Schaffer collateral stimulation. MAM treatment did not affect presynaptic neurotransmitter release as measured by paired-pulse facilitation ($66 \pm 8\%$ increase for saline and $57 \pm 11\%$ for MAM-treated rats) ($P > 0.05$) or paired-pulse inhibition ($30 \pm 12\%$ decrease for saline and $35 \pm 7\%$ for MAM treated-rats) ($P > 0.05$). Treatment did not affect cell excitability as measured by changes in the input-output curve ($P > 0.05$) (Fig. 3a). Trained groups treated with MAM or saline showed long-term potentiation ($F(14, 140) = 5.46$, $P < 0.00001$) and the degree of potentiation was not different between groups ($P > 0.05$) (Fig. 3b). Because field potential measurements cannot detect a loss or silencing of mature cells, these measures reflect physiological responsiveness and cell excitability for those that remain. However, these results suggest that MAM treatment does not alter gross neurophysiological responsiveness in the hippocampus despite its detrimental effect on trace conditioning.

In a third set of experiments, we evaluated whether the detrimental effect of MAM on trace conditioning was due to general aversive effects of the drug on performance. We also identified the time period when the new cells become involved in learning. Groups of rats were treated with MAM (5 mg kg^{-1} , $n = 6$) or saline ($n = 6$) for 6 d to roughly ascertain a time when new cells were too young to participate in learning but mature cells could be affected by the drug. They were injected with BrdU on days 1, 3 and 5. Additional groups were treated as before with MAM ($n = 3$) or saline ($n = 3$) for 14 d (injected with BrdU on days 10, 12 and 14). All groups were

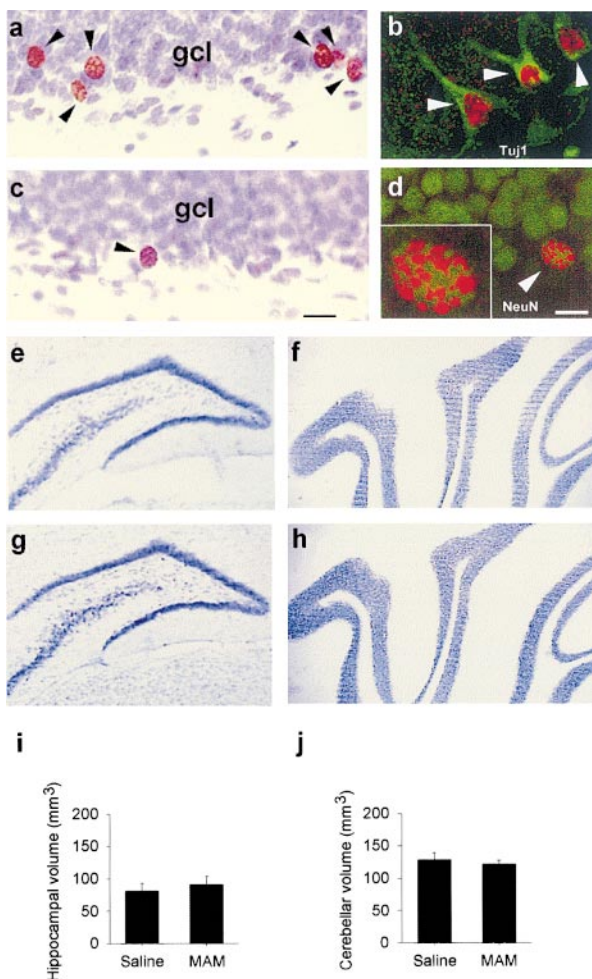


Figure 2 MAM treatment decreases number of BrdU-labelled cells in dentate gyrus and does not alter gross morphology. **a**, BrdU-labelled cells in saline-treated rat (arrowheads). gcl, granule cell layer. **b**, Cells labelled with BrdU (red, nuclei) and TuJ1 (green, cytoplasm). **c**, BrdU-labelled cells in MAM-treated rat (arrowhead). **d**, Cell labelled with BrdU (red, nucleus) and NeuN (green, nucleus). Insert depicts a cell stained with NeuN. **e**, **f**, Cresyl violet stained section of dentate gyrus (**e**) and cerebellar cortex (**f**) from a saline-treated rat. **g**, **h**, Cresyl violet stained section of dentate gyrus (**g**) and cerebellar cortex (**h**) from a MAM-treated rat. **i**, **j**, Hippocampal (**i**) and cerebellar (**j**) volume (mean $\pm$ s.e.m.). Scale bars: **a**, **c**, 30 μm ; **b**, **d**, 10 μm .

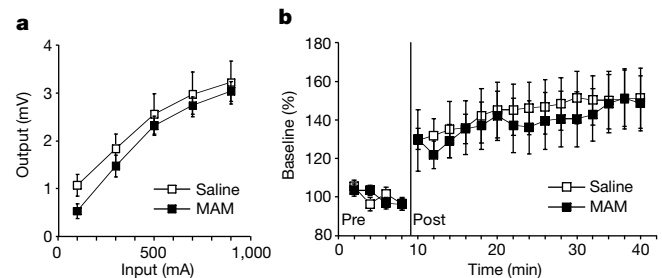


Figure 3 Reduction of adult-generated granule cells in the hippocampus does not affect neurophysiological responses in area CA1. **a**, Input-output curves for groups treated with MAM versus saline. **b**, Long-term potentiation of EPSP in area CA1 of the hippocampus in response to theta burst stimulation of Schaffer collaterals. Post-tetanic values (post) are expressed as a percentage of the pretetanus baseline (pre).

trace conditioned for 4 d, beginning 2 d after treatment. Six days of MAM treatment did not alter the number of conditioned responses ($F(1,10) = 0.75$, $P > 0.05$) or acquisition (an interaction between conditioned responses and trials of training) ($F(11,110) = 1.61$, $P > 0.05$) (Fig. 4b) despite a 78% decrease in number of BrdU-labelled cells (2,380 for MAM compared with 11,000 for saline) ($F(1,10) = 67.58$, $P < 0.000001$) (Fig. 4a). In contrast, 14 d of MAM treatment impaired the acquisition of trace conditioning ($F(11,44) = 10.04$, $P < 0.0000001$) (Fig. 4d) and reduced the numbers of new cells by 83% (1,472 for MAM compared with 9,144 for saline) ($F(1,4) = 15.11$, $P < 0.02$) (Fig. 4c). Thus, treatment for 6 d did not impair trace conditioning whereas treatment for 14 d did.

MAM treatment did not alter levels of motor activity in rats treated for 6 d ($P > 0.05$) or 14 d ($P > 0.05$), nor did it affect pain sensitivity (6 d groups, $P > 0.05$; 14 d groups, $P > 0.05$). MAM treatment did not alter blood levels of the stress hormone corticosterone (mean $\pm$ s.e.m.) in those treated for 6 d ($P > 0.05$) (124 ± 24 ng ml $^{-1}$ for MAM and 121 ± 13 ng ml $^{-1}$ for saline) or 14 d ($P > 0.05$) (101 ± 31 ng ml $^{-1}$ for MAM and 92 ± 7 ng ml $^{-1}$ for saline). These results suggest that treatment *per se* does not alter performance, nor does it affect general activity, pain sensitivity or stress levels, as indicated by endogenous levels of glucocorticoids. Moreover, these results indicate that the new cells are about 1–2 weeks of age when they become involved in the learned response.

We conducted stereological analyses on the hippocampus and cerebellum to determine whether the MAM treatment caused structural changes other than a reduction in the number of new cells (Fig. 2). MAM treatment did not alter the volume of the hippocampal formation ($P > 0.05$) (Fig. 2e, g, i) nor the cerebellum ($P > 0.05$) (Fig. 2f, h, j). MAM treatment did not alter the volume of the granule ($P > 0.05$) or CA1 pyramidal cell layer ($P > 0.05$), nor did it reduce the number of mature granule neurons in the dentate

gyrus ($P > 0.05$). There was no evidence of degenerating cells in these regions.

We next determined whether trace conditioning would recover as neurogenesis did. Groups were treated with MAM ($n = 5$) or saline ($n = 5$) for 14 d, followed by 21 d without treatment, then trace conditioned for 4 d and perfused 24 h later. They were injected with BrdU 10, 8 and 6 d before perfusion. At the end of the 21-day recovery period, the number of BrdU-labelled cells was not different in groups treated with MAM (6,064) versus saline (7,200) ($P > 0.05$) (Fig. 4e). The ability to acquire the trace conditioned response recovered, and there was no difference between groups ($F(11,88) = 0.77$, $P > 0.05$) (Fig. 4f). Thus, replenishment of immature neurons after MAM treatment was associated with efficient acquisition of trace memories.

The survival of adult-generated dentate granule neurons is enhanced in response to hippocampal-dependent but not hippocampal-independent learning¹³. Specifically, survival of cells generated between one and two weeks earlier was enhanced in response to trace conditioning and spatial navigation learning, but not delay conditioning or cue-directed navigation learning. Our data indicate that these cells are not only affected by, but involved in, hippocampal-dependent learning, at least in trace conditioning. As rats treated with MAM for 6 d were not impaired, whereas those treated for 14 d were, the results are consistent with the idea that the cells become critical about 1–2 weeks after their generation. By this time, newly generated cells have become incorporated into the granule cell layer, form dendrites and extend axons into the CA3 region^{21,26}. This ability to undergo rapid structural change may be a characteristic of immature neurons that makes them ideally suited for forming associations between stimuli.

It has been suggested that hippocampal-dependent learning processes, such as those involved in trace conditioning, involve conscious awareness or recollection, and that the hippocampus performs such functions^{27,28}. On the basis of the relative immaturity

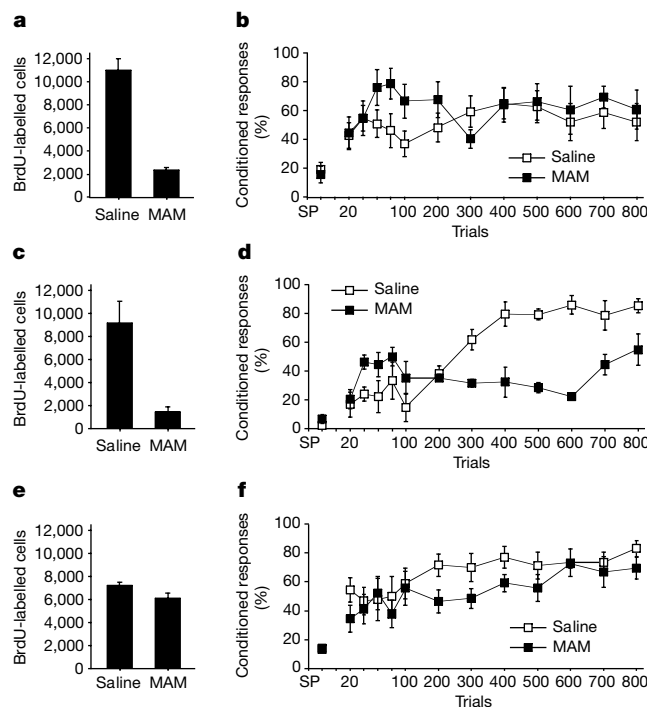


Figure 4 Trace conditioning is not affected when newly generated cells are less than one week of age or proliferation recovers. **a**, Total numbers of BrdU-labelled cells in dentate gyrus when treated with MAM or saline for 6 d. (mean $\pm$ s.e.m.). **b**, SP and percentage of conditioned responses during trace conditioning. **c**, Total numbers of BrdU-labelled cells in dentate gyrus when treated with MAM or saline for 14 d. (mean $\pm$ s.e.m.). **d**, SP and

percentage of conditioned responses during trace conditioning. **e**, Total numbers of BrdU-labelled cells in the dentate gyrus when treated with MAM or saline for 14 d. Training began 3 weeks later. **f**, SP and percentage of conditioned responses during trace conditioning.

of the neurons and their apparent sensitivity to both temporal and spatial associations, it could be hypothesized that the cells are used to maintain stimulus representations of newly experienced stimuli in the event that these representations are associated with subsequent events in time and space^{29,30}. Although it seems that the new cells are associated with the formation of hippocampal-dependent memories, we note that rats treated with saline learned the delay task faster than trace. Thus, newly generated neurons may not be used for learning under more lenient conditions, but become involved as task demands increase.

A reduction in the number of newly generated neurons in the dentate gyrus of the hippocampal formation impaired hippocampal-dependent, but not hippocampal-independent, forms of associative memory formation. These results suggest that immature neurons in the adult brain are involved and perhaps necessary for the acquisition of new hippocampal-dependent memories about temporal relations or the accurate timing of learned responses, such as during the acquisition of trace memories. □

Methods

Classical eyeblink conditioning

We connected headstages to four implanted electrodes: two delivered the unconditioned stimulus and two transmitted eyelid EMG, which was filtered to pass 0.3–1.0 kHz and amplified (10K). Rats were acclimated to the conditioning apparatus for 1 h, and spontaneous blink rate was recorded. Twenty-four hours later, rats were exposed to ten white noise stimuli (320 ms, 83 dB), and eyeblinks during the first 100 ms of the white noise stimulus were considered sensitized responses. We exposed rats to 200 paired trials a day for 4 d, for a total of 800 trials. They were exposed to either a delay model in which an 850 ms, 83 dB burst of white noise conditioned stimulus overlapped and co-terminated with a 100 ms, 7 mA periorbital shock unconditioned stimulus, or a trace model in which a 250 ms conditioned stimulus was separated from the unconditioned stimulus by 500 ms. Interstimulus interval was 750 ms for both of the models. The intertrial interval was 20 ± 10 ms. After training, we perfused rats and stained the brains for detection of BrdU and neuronal markers TuJ1 and NeuN.

Maximum EMG response that occurred during a 250 ms pre-stimulus baseline was added to four times its standard deviation, and responses that exceeded that value and had a width greater than 3 ms were considered eyeblinks. Those that occurred within 500 ms of unconditioned stimulus onset were considered conditioned responses. For data analysis we used analysis of variance (ANOVA) with repeated measures and planned comparisons.

Activity, pain sensitivity and stress hormones

For motor activity, horizontal and vertical motions were recorded during 30 min of free movement in a 30 cm × 30 cm chamber. For pain analysis, rats were gently restrained and tails were exposed to a beam of light of increasing intensity. Latency to withdraw was measured four times and averaged. We collected, centrifuged and stored the trunk blood for radio-immunoassay of corticosterone (Coat-a-Count).

Neurophysiological recording procedures

Within 5 d of training, subgroups were anaesthetized and placed in a stereotaxic instrument. A concentric bipolar stimulating electrode was placed 3.5 mm posterior and 3.5 mm lateral to Bregma, and a recording electrode was positioned 3.5 mm posterior and 2.7 mm lateral to Bregma. Responses were amplified 1K, filtered (1 Hz–1 kHz) and collected using BrainWave software. Input–output curves were generated by manipulating current stimulation between V_{min} and V_{max} , averaging six responses at five intensities. Maximum response was reduced to half and data from paired-pulse facilitation (40 ms interstimulus interval) and inhibition (20 ms interstimulus interval) were collected. Baseline was established as 1/2 V_{max} and a tetanus of ten, 40 ms 100 Hz bursts at 5 Hz was delivered. Responses were collected every 30 s, averaging every four, for 30 min, and are presented as a percentage of pretetanus baseline. We perfused rats and later stained the cells for detection of BrdU and neuronal markers, TuJ1 and NeuN.

Histology

Brains were cut on an oscillating tissue slicer (40-μm thick sections) throughout the dentate gyrus and cerebellum. We stained tissue for BrdU using peroxidase and fluorescence techniques. For peroxidase, tissue was heated at 75 °C for 1 h, followed by 2 M HCl for 1 h. After rinsing in PBS, sections were incubated overnight at 4 °C in mouse monoclonal anti-BrdU (1:250; Novocastra). We reacted the sections using a standard Vector ABC kit. For immunofluorescence, sections were treated with 2 M HCl for 30 min. After rinsing in TBS, tissue was incubated in rat monoclonal anti-BrdU (1:250; Accurate) and then in mouse monoclonal anti-NeuN (1:200; Chemicon), mouse monoclonal anti-TuJ1 (1:500, a gift from A. Frankfurter), or rabbit polyclonal anti-TUC-4 (antibody 25, TOAD/ulip/crm-4, formerly TOAD-64; 1:10,000, a gift from S. Hockfield). Sections were reacted with biotinylated anti-rat followed by streptavidin-Alexa 568 and anti-mouse or anti-rabbit Alexa 488.

We coded slides before quantitative analysis. Stereological analyses of total number of BrdU-labelled cells in the dentate gyrus were conducted using a modified version of the optical fractionator method on every twelfth section of peroxidase stained tissue. For each section, the number of BrdU-labelled cells was determined in the dentate gyrus, excluding those in the outermost focal plane to avoid counting cell caps. Resulting numbers were tallied and multiplied by tissue thickness (40 μm) and number of intervening sections ($n = 12$). Double labelling of BrdU and NeuN or BrdU and TuJ1 was assessed with a fluorescence microscope (Olympus BX-60) and a Zeiss confocal laser scanning microscope (510 LSM; lasers HENE 543 NM and ARGON 458/488 NM). Estimates of the total number of mature granule cells were carried out using an optical fractionator method (Stereoinvestigator, Microbrightfield). Volume estimates of the cerebellum and hippocampus were determined using Cavalieri's principle, with area measurements obtained from every twelfth section using Stereoinvestigator.

Received 16 October; accepted 28 November 2000.

- Altman, J. & Das, G. D. Autoradiographic and histological evidence of postnatal hippocampal neurogenesis in rats. *J. Comp. Neurol.* **124**, 319–335 (1965).
- Kaplan, M. S. & Hinds, J. W. Neurogenesis in the adult rat: electron microscopic analysis of light radioautographs. *Science* **197**, 1092–1094 (1977).
- Barnea, A. & Nottebohm, F. Seasonal recruitment of hippocampal neurons in adult free-ranging black-capped chickadees. *Proc. Natl Acad. Sci. USA* **91**, 11217–11221 (1994).
- Alvarez-Buylla, A. & Nottebohm, F. Migration of young neurons in adult avian brain. *Nature* **335**, 353–354 (1988).
- Cameron, H. & McKay, R. D. Restoring production of hippocampal neurons in old age. *Nature Neuroscience* **2**, 894–897 (1999).
- Eriksson, P. S. et al. Neurogenesis in the adult human hippocampus. *Nature Med.* **4**, 1313–1317 (1998).
- Gould, E., Reeves, A., Graziano, M. S. A. & Gross, C. Neurogenesis in the neocortex of adult primates. *Science* **286**, 548–552 (1999).
- Gould, E., Tanapat, P., McEwen, B., Flugge, G. & Fuchs, E. Proliferation of granule cell precursors in the dentate gyrus of adult monkeys is diminished by stress. *Proc. Natl Acad. Sci. USA* **95**, 3168–3171 (1998).
- Wagner, J. P., Black, I. B. & DiCicco-Bloom, E. Stimulation of neonatal and adult brain neurogenesis by subcutaneous injection of basic fibroblast growth factor. *J. Neurosci.* **19**, 6006–6016 (1999).
- Lopez-Garcia, C., Molowny, A., Garcia-Verdugo, J. M. & Ferrer, I. Delayed postnatal neurogenesis in the cerebral cortex of lizards. *Brain Res.* **471**, 167–174 (1988).
- Lois, C. & Alvarez-Buylla, A. Long-distance neuronal migration in the adult mammalian brain. *Science* **264**, 1145–1148 (1994).
- Nottebohm, F. From bird song to neurogenesis. *Sci. Am.* **260**, 74–79 (1989).
- Gould, E., Beylin, A. V., Tanapat, P., Reeves, A. & Shors, T. J. Learning enhances adult neurogenesis in the hippocampal formation. *Nature Neuroscience* **2**, 260–265 (1999).
- Gould, E., Tanapat, P., Hastings, N. & Shors, T. J. Neurogenesis in adulthood: a possible role in learning. *Trends Cog. Neurosci.* **3**, 186–192 (1999).
- Solomon, P. R., van der Schaaf, E. R., Weisz, D. & Thompson, R. F. Hippocampus and trace conditioning of the rabbit's classically conditioned nictitating membrane response. *Neuroscience* **100**, 729–744 (1986).
- Schmaltz, L. W. & Theios, J. Acquisition and extinction of a classically conditioned response in hippocampectomized rabbits (*Oryctolagus cuniculus*). *J. Comp. Physiol. Psychol.* **79**, 328–333 (1972).
- Berger, T. W. & Orr, W. B. Hippocampectomy selectively disrupts discrimination reversal conditioning of the rabbit nictitating membrane response. *Behav. Brain Res.* **8**, 49–68 (1983).
- Johnston, M. V. & Coyle, J. T. Histological and neurochemical effects of fetal treatment with methylazoxymethanol on rat neocortex in adulthood. *Brain Res.* **170**, 135–155 (1979).
- Wood, K. A. & Yule, R. J. The role of free radicals and p53 in neuron apoptosis *in vivo*. *Neuroscience* **15**, 5851–5857 (1995).
- Minger, S. L. & Davies, P. Persistent innervation of the rat neocortex by basal forebrain cholinergic neurons despite the massive reduction of cortical target neurons. Morphometric analysis. *Exp. Neurol.* **117**, 124–138 (1992).
- Hastings, N. B. & Gould, E. Rapid extension of axons into the CA3 region by adult-generated granule cells. *J. Comp. Neurol.* **413**, 146–154 (1999).
- Miller, M. W. & Nowakowski, R. S. Use of bromodeoxyuridine-immunohistochemistry to examine the proliferation, migration and time of origin of cells in the central nervous system. *Brain Res.* **457**, 44–52 (1988).
- Quinn, C. C., Gray, G. E. & Hockfield, S. A family of proteins implicated in axon guidance and outgrowth. *J. Neurobiol.* **41**, 158–164 (1999).
- Weiss, C., Bouwmeester, H., Power, J. M. & Disterhoft, J. F. Hippocampal lesions prevent trace eyeblink conditioning in the freely moving rat. *Behav. Neurosci.* **104**, 243–252 (1990).
- McEchron, M. D., Bouwmeester, H., Tseng, W., Weiss, C. & Disterhoft, J. F. Hippocampectomy disrupts auditory trace fear conditioning and contextual fear conditioning in the rat. *Hippocampus* **8**, 638–646 (1998).
- Tanapat, P., Hastings, N. B., Reeves, A. J. & Gould, E. Estrogen stimulates a transient increase in the number of new neurons in dentate gyrus of the adult female rat. *J. Neurosci.* **19**, 5792–5801 (1999).
- Clark, R. E. & Squire, L. R. Classical conditioning and brain systems: the role of awareness. *Science* **280**, 77–81 (1998).
- LaBar, K. S. & Disterhoft, J. F. Conditioning, awareness, and the hippocampus. *Hippocampus* **8**, 620–626 (1998).
- Wallenstein, G., Eichenbaum, H. & Hasselmo, M. The hippocampus as an associator of discontiguous events. *Trends Neurosci.* **21**, 317–323 (1998).
- Shors, T. J., Beylin, A. V., Wood, G. E. & Gould, E. The modulation of Pavlovian memory. *Behav. Brain Res.* **110**, 39–52 (2000).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank L. D. Matzel for comments and L. King, G. Kozorovitskiy and B. DiCicco-Bloom for technical assistance. This work is supported by NIMH, NSF, the National Alliance for Research on Schizophrenia and Depression, and the van Ameringen Foundation to T.J.S., and NIMH to E.G.

Correspondence and requests for materials should be addressed to T.J.S. (e-mail: shors@rci.rutgers.edu).

Effects of chronic exposure to cocaine are regulated by the neuronal protein Cdk5

James A. Bibb^{*}, Jingshan Chen[†], Jane R. Taylor[†], Per Svenningsson^{*}, Akinori Nishi[‡], Gretchen L. Snyder^{*}, Zhen Yan^{*§}, Zachary K. Sagawa^{*}, Charles C. Ouimet^{||}, Angus C. Nairn^{*}, Eric J. Nestler^{†¶} & Paul Greengard^{*}

^{*} Laboratory of Molecular and Cellular Neuroscience, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA

[†] Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut 06520, USA

[‡] Department of Physiology, Kurume University School of Medicine, Kurume, Fukuoka 830-0011, Japan

^{||} Program in Neuroscience, Florida State University, Tallahassee, Florida 32306, USA

[¶] Department of Psychiatry, The University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA

Cocaine enhances dopamine-mediated neurotransmission by blocking dopamine re-uptake at axon terminals. Most dopamine-containing nerve terminals innervate medium spiny neurons in the striatum of the brain. Cocaine addiction is thought to stem, in part, from neural adaptations that act to maintain equilibrium by countering the effects of repeated drug administration^{1,2}. Chronic exposure to cocaine upregulates several transcription factors that alter gene expression and which could mediate such compensatory neural and behavioural changes^{1–4}. One such transcription factor is Δ FosB, a protein that persists in striatum long after the end of cocaine exposure^{3,5}. Here we identify *cyclin-dependent kinase 5* (*Cdk5*) as a downstream target gene of Δ FosB by use of DNA array analysis of striatal material from inducible transgenic mice. Overexpression of Δ FosB, or chronic cocaine administration, raised levels of Cdk5 messenger RNA, protein, and activity in the striatum. Moreover, injection of Cdk5 inhibitors into the striatum potentiated behavioural effects of repeated cocaine administration. Our results suggest that changes in Cdk5 levels mediated by Δ FosB, and resulting alterations in signalling involving D1 dopamine receptors, contribute to adaptive changes in the brain related to cocaine addiction.

Transgenic mice displaying inducible and targeted expression of Δ FosB in the nucleus accumbens and caudatoputamen—together, the striatum of the brain—were engineered⁶. Analysis of complementary DNA expression array profiles from mice overexpressing Δ FosB indicated that the neuronal protein kinase Cdk5 was a downstream target gene for Δ FosB in these brain regions (Fig. 1a). This effect was confirmed by quantitative *in situ* hybridization analyses of coronal brain sections from mice that did (H₂O) or did not (Dox) overexpress Δ FosB (Fig. 1b, left). Increased expres-

sion of Cdk5 mRNA in response to Δ FosB accumulation was evident in both the caudatoputamen ($160.0 \pm 14.2\%$ of control, $P < 0.05$) and the nucleus accumbens ($152.6 \pm 12.9\%$, $P < 0.05$). Adult rats injected with cocaine for 8 days showed elevated levels of Cdk5 mRNA in the caudatoputamen ($151.8 \pm 11.8\%$, $P < 0.05$) and nucleus accumbens ($150.5 \pm 10.0\%$, $P < 0.05$) in comparison to animals injected with saline (Fig. 1b, left). An increase was also observed in the level of mRNA encoding the neuron-specific Cdk5-activating cofactor, p35, in response to Δ FosB induction or chronic cocaine administration in both the caudatoputamen ($121.8 \pm 5.7\%$ and $129.8 \pm 7.5\%$, respectively, $n = 6$, $P < 0.05$) and the nucleus accumbens ($119.8 \pm 5.2\%$ and $126.3 \pm 7.4\%$, respectively, $n = 6$, $P < 0.05$) (Fig. 1b, right). Increased levels of Cdk5 protein were observed in striatal tissue dissected from transgenic Δ FosB-expressing mice ($130 \pm 10\%$, $P < 0.05$) and from rats exposed chronically to cocaine ($180 \pm 20\%$, $P < 0.05$) compared to control animals (Fig. 1c, left and middle panels). Increased striatal p35 protein levels also occurred in response to chronic cocaine ($147.3 \pm 11\%$, $n = 12$, $P < 0.05$) (Fig. 1c, right panel). These results indicated that Cdk5, together with its activating cofactor p35, are downstream targets of chronic cocaine exposure, and raised the possibility that this protein kinase is involved in the behavioural effects of cocaine.

A characteristic behavioural effect of cocaine is potentiation of locomotor activity. The effects of daily intra-accumbens infusions of a potent Cdk5 inhibitor, roscovitine, on cocaine-induced locomotor activity were examined over a 60-min period daily for 5 days. Administration of cocaine resulted in a marked and progressive increase in locomotor activity over this 5-day period, indicating the development of locomotor sensitization, during the initial 10–20 min period following injection (data not shown). Roscovitine infusions did not significantly affect locomotor responses to initial cocaine administration. However, roscovitine markedly potentiated the locomotor effects of repeated cocaine exposures. This was evident as an augmentation of cocaine's effects over successive days of injections (Fig. 2a). By day 4, significant differences were observed between the saline/cocaine and roscovitine/cocaine groups. By day 5, mean cocaine-induced activity rates for roscovitine-infused animals were almost double those measured for vehicle-infused animals. This effect of roscovitine was most evident 40–60 min after cocaine administration (Fig. 2b).

Repeated intra-accumbens infusions of a less selective Cdk5 inhibitor, olomoucine, also potentiated cocaine's locomotor effects. This action was similar to that produced by roscovitine, except that a marked behavioural effect was already observed as early as day 3 (Fig. 2c). In response to cocaine treatment on days 4 and 5, olomoucine-treated animals exhibited stereotypy, which denotes enhanced sensitization and is known to compete with increases in locomotor activity¹⁰. In contrast, intra-accumbens infusions of the inactive congener, iso-olomoucine, failed to enhance either locomotor (Fig. 2c) or stereotypic responses to cocaine. These behavioural findings indicate that cocaine-induced increases in Cdk5 levels may serve a homeostatic function to dampen responses to subsequent drug exposure.

One way in which Cdk5 could regulate the psychomotor effects of chronic cocaine is through regulation of dopamine signalling. Cdk5 phosphorylates a key molecule involved in striatal dopamine signalling, DARPP-32—dopamine and cyclic AMP-regulated phosphoprotein, M_r 32K—at threonine-75 (Thr 75)⁷. Immunocytochemistry studies revealed that DARPP-32 and Cdk5 are colocalized: all medium spiny neurons in the nucleus accumbens that contained DARPP-32 also contained Cdk5 (Fig. 3a). Furthermore, dendrites in the neuropil were typically double-labelled and sometimes contained puncta that were immunoreactive to both proteins. Increased levels of phospho-Thr 75 DARPP-32 were observed in striatal tissue dissected from transgenic mice overexpressing Δ FosB (Fig. 3b). In rats subjected to chronic cocaine

[§] Present address: Department of Physiology and Biophysics, School of Medicine and Biomedical Sciences, State University of New York at Buffalo, Buffalo, New York 14214, USA.

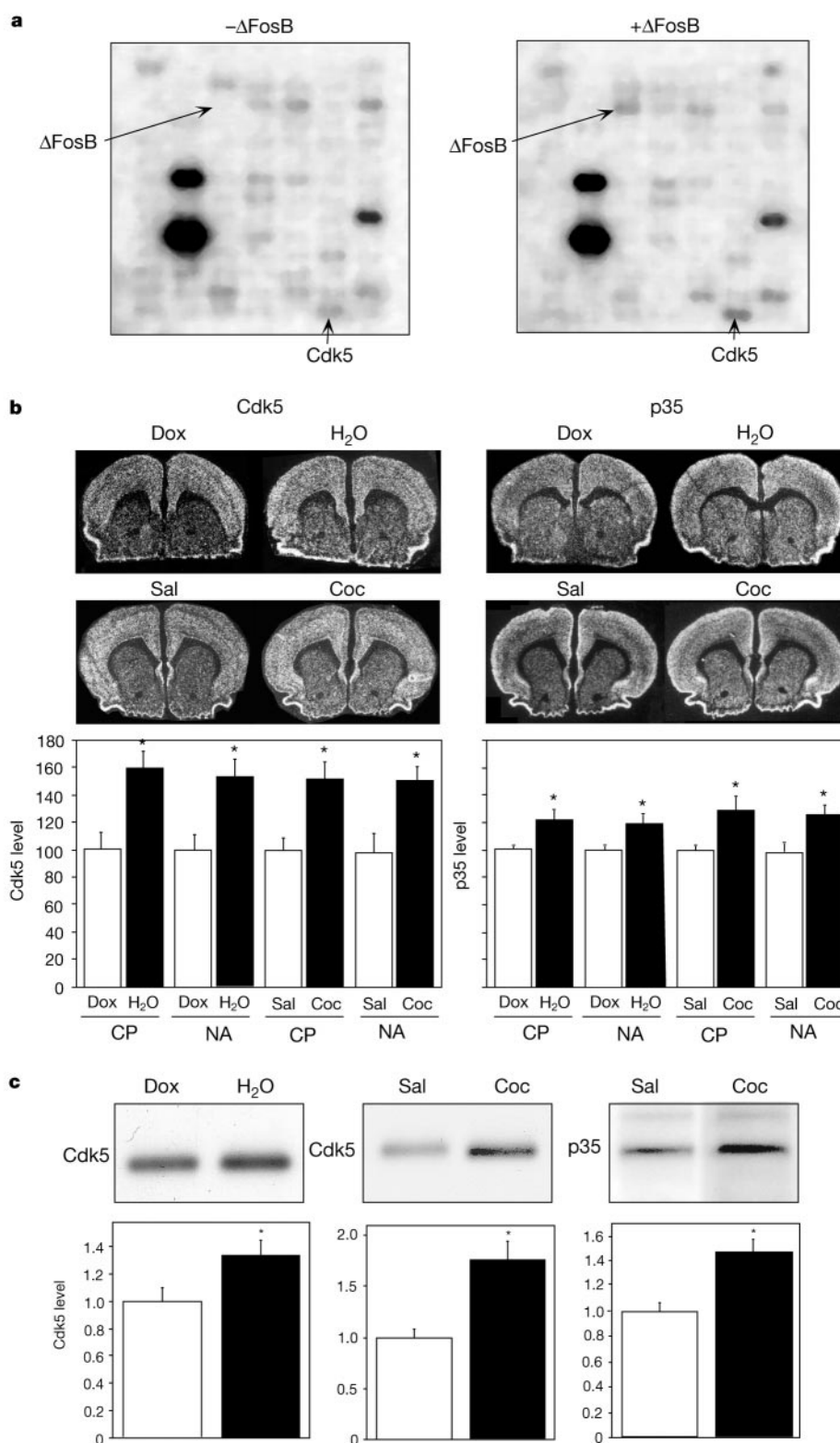


Figure 1 Increased expression of Cdk5 in inducible transgenic mice overexpressing Δ FosB and in rats chronically treated with cocaine. **a**, cDNA expression arrays probed with radiolabelled cDNA from control mice ($-\Delta$ FosB) (left) or transgenic mice overexpressing Δ FosB ($+\Delta$ FosB) (right). The positions of oligonucleotides encoding Δ FosB and Cdk5 are indicated. **b**, Comparison of the levels of Cdk5 (left) and p35 (right) gene expression by *in situ* hybridization. Representative labellings are shown for sections from inducible transgenic mice on (Dox) or off doxycycline (H₂O) (top panels) and rats treated chronically with saline (Sal) or cocaine (Coc; middle panels). Levels of signals in

the caudatoputamen (CP) and nucleus accumbens (NA) are shown in bar graphs. **c**, Comparison of the levels of Cdk5 and p35 protein. Representative Cdk5 and p35 immunoblots of striatal tissue dissected from inducible transgenic mice on or off doxycycline (left) and from rats treated chronically with saline or cocaine (middle, right) are shown with levels. Data represent means $\pm$ s.e.m. for $n = 6$; asterisk indicates $P < 0.05$ compared to control, Student's *t*-test. Data for Cdk5 and p35 levels are expressed in arbitrary units.

administration, the levels of phospho-Thr 75 DARPP-32 were increased in both the caudatoputamen and nucleus accumbens (Fig. 3c, d). Total levels of DARPP-32 were unaffected under these conditions.

Cdk5-dependent phosphorylation of DARPP-32 at Thr 75 reduces the efficacy of dopamine/cyclic AMP/protein kinase A/DARPP-32/protein phosphatase-1 signalling⁷. Various aspects of this signalling pathway were examined. Treatment of striatal slices

from saline-injected rats with the D1 receptor agonist, SKF 81297, is known to increase PKA-dependent phosphorylation of DARPP-32 (ref. 8), ARPP-16 (ref. 9) and ARPP-21 (ref. 10). These effects of the D1 agonist were attenuated in striatal slices from rats exposed to chronic cocaine (Fig. 4a). Decreased PKA-dependent phosphorylation of DARPP-32 (at Thr 34) and of the GluR1 subunit of the AMPA-type glutamate receptor (at Ser 845) was observed in striatal tissue dissected from rats chronically treated with cocaine (Fig. 4b). Consistent with these effects being under the control of Cdk5, roscovitine treatment restored PKA-dependent phosphorylation of DARPP-32 at Thr 34 (10.6 ± 2.1 -fold increase, $n = 6$, $P = 0.0002$) and the GluR1 receptor at Ser 845 (5.2 ± 1.6 -fold increase, $n = 6$, $P = 0.006$) in striatal slices from chronic-cocaine-treated rats. Consistent with the ability of PKA-dependent phosphorylation of GluR1 to increase the conductance of AMPA/kainate channels¹¹, we found reduced peak amplitudes of AMPA/kainate-evoked currents in striatal neurons from chronic-cocaine-treated rats (Fig. 4c). These results expand upon the observation that nucleus accumbens neurons from rats treated with a psychomotor stimulant were less sensitive to the current-dependent increases in firing rates of striatal dopaminergic neurons caused by glutamate¹².

It has been suggested that chronic administration of cocaine, as well as other drugs of abuse, lead to enhancement of D1/PKA

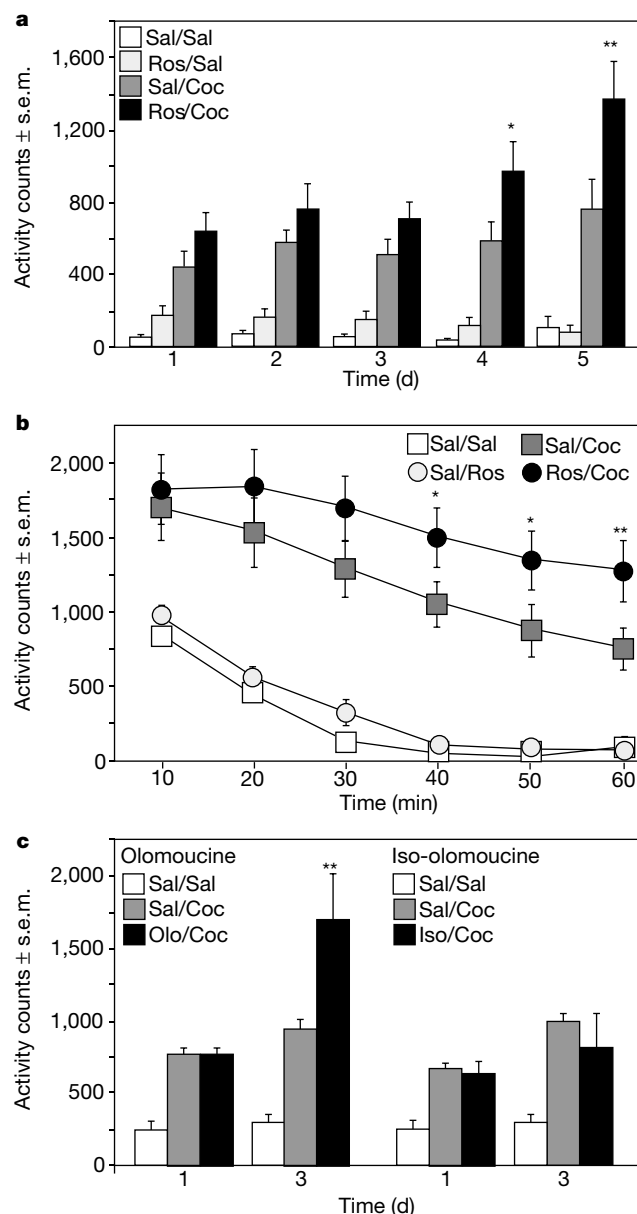


Figure 2 Effect of Cdk5 inhibitors on locomotor behavioural response to repeated cocaine injections. **a**, Locomotor activity for rats, given intra-accumbens infusions of saline (Sal) or roscovitine (Ros) and i.p. injections of saline (Sal) or cocaine (Coc), from 50 to 60 min post-injection on successive days 1–5. **b**, Behavioural effects of each treatment on day 5 measured for 60 min post-injection with values plotted at 10-min intervals. **c**, Behavioural effects of intra-accumbens infusion of olomoucine (Olo; left panel) or the inactive analogue, iso-olomoucine (Iso; right) on cocaine (i.p.)-induced locomotor activity from 40–50 min post-injection at days 1 and 3 of treatment. All data represent mean activity counts $\pm$ s.e.m. measured by photocell over test period; asterisk indicates $P < 0.05$, double asterisk indicates $P < 0.001$ by analysis of variance and *post hoc* Scheffé's *F*-test. For day 3 in panel **a**, $P < 0.07$. For panels **a** and **b**, Sal/Sal, $n = 6$; Sal/Coc, $n = 11$; Ros/Sal, $n = 9$; Ros/Coc, $n = 11$. For panel **c**, Sal/Sal, $n = 6$, Sal/Coc, $n = 9$, Olo/Coc, $n = 11$; Iso/Coc, $n = 11$.

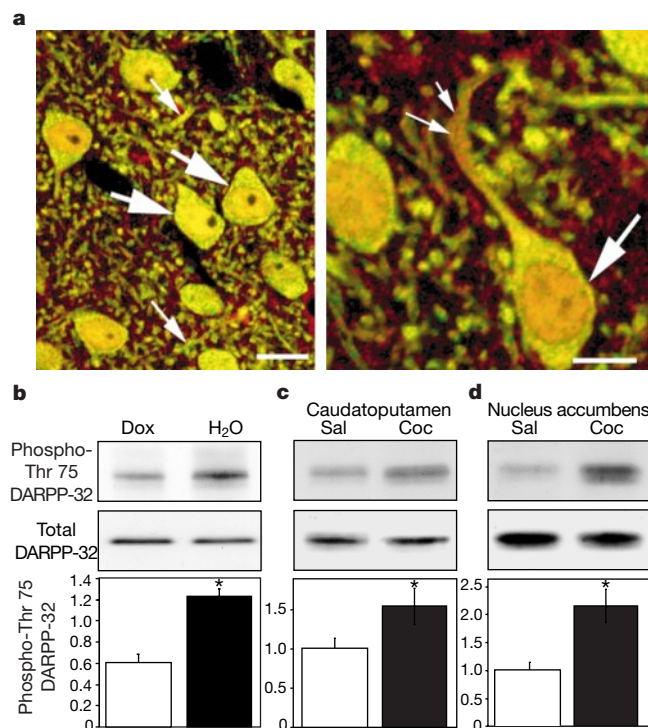


Figure 3 Increased phosphorylation of DARPP-32 by Cdk5 in inducible transgenic mice overexpressing Δ FosB and in rats chronically treated with cocaine. **a**, Laser confocal photomicrograph of tissue co-immunolabelled for Cdk5 (red) and DARPP-32 (green) in the nucleus accumbens. Overlapping signal appears prominently as yellow. All neurons that immunostained for DARPP-32 also immunostained for Cdk5 (see larger arrows for examples). Dendrites in the neuropil were also double-stained (small arrows). Images were acquired separately in each channel (dual scan mode) to eliminate the possibility of signal bleed-over from one channel to the other. Scale bar, 10 μ m. **b**, Quantitative immunoblot analysis of level of phospho-Thr 75 DARPP-32 in striatal tissue dissected from inducible Δ FosB transgenic mice on or off doxycycline. Representative blots are shown for phospho-Thr 75 DARPP-32 and total DARPP-32 in the upper two panels and levels are shown in the lower panel. **c**, **d**, Level of phospho-Thr 75 DARPP-32 in caudatoputamen (**c**) and nucleus accumbens (**d**) from rats treated with saline or chronic cocaine for 10 d. Data represent means $\pm$ s.e.m. for $n = 6$; asterisk indicates $P < 0.05$, Mann-Whitney non-parametric *t*-test.

signalling^{2,13}. However, our results indicate that at least some target substrates exhibit reduced D1/PKA-dependent phosphorylation following chronic exposure to cocaine. Our observations of increased Cdk5 in response to cocaine, which can cause an increase in phospho-Thr 75 DARPP-32 and a reduction in PKA activity, provides a mechanism for differential regulation of PKA-dependent phosphorylation by chronic cocaine. It is possible that the Δ FosB/Cdk5/phospho-Thr 75 DARPP-32 pathway is targeted towards particular substrates or subcellular elements and away from others. Moreover, the effects of chronic cocaine and other drugs of abuse on dopamine signalling may be subject to variations in dosing parameters, period of withdrawal, and paradigms of study.

There is now considerable evidence that the D1 receptor is important in mediating the behavioural effects of cocaine^{12,14–16}. We have reported that an acute dose of cocaine activates PKA, causing an increase in phosphorylation of DARPP-32 at Thr 34 while causing a decrease in phosphorylation at Thr 75 via a signalling mechanism that depends on protein phosphatase 2A¹⁷. The data

reported here support a distinct biochemical scheme in which repeated exposure to cocaine causes accumulation of Δ FosB, which in turn results in increased expression of Cdk5; increased Cdk5-dependent phosphorylation of DARPP-32 at Thr 75 then attenuates D1/PKA signalling. In support of the behavioural relevance of this scheme, mice lacking the *DARPP-32* gene¹⁸ and rats treated with Cdk5 inhibitors (Fig. 2) both exhibit enhanced behavioural responses to chronic administration of cocaine.

Previous studies established that induction of Δ FosB expression in the striatum enhances behavioural responsiveness to cocaine⁴. Thus Δ FosB-mediated induction of Cdk5 would appear to oppose the overall effect of this transcription factor, suggesting that Δ FosB—which acts on numerous genes—mediates both sensitizing and compensatory adaptations. The ability of intra-accumbens infusions of Cdk5 inhibitors to potentiate the locomotor effects of repeated cocaine administration suggests that the Cdk5/DARPP-32 pathway plays a negative feedback homeostatic role with respect to the behavioural effects of cocaine. It remains to be determined whether other actions of Cdk5^{19–21} contribute to the behavioural effects of cocaine. Nevertheless, by combining DNA array analyses, an inducible transgenic animal model, and behavioural studies, we have now identified Cdk5 as a target that alters dopamine signal transduction as part of the long-term plasticity associated with cocaine addiction. □

Methods

For cDNA expression arrays, total RNA was isolated with the RNAqueous phenol-free total RNA isolation kit (Ambion) from dissected striatal tissue from either control mice carrying only the *NSE-tTA* gene ($-\Delta$ FosB) or Δ FosB-inducible transgenic mice carrying both *Tetop- Δ FosB* and *NSE-tTA* genes, which had been fed 100 mg l⁻¹ doxycycline to inhibit transgene expression, followed by 12 weeks in the absence of doxycycline ($+\Delta$ FosB). Poly(A)⁺ RNA was isolated from total RNA using an Oligotex mRNA isolation kit (Qiagen), and used as template for synthesis of ³²P-labelled cDNA probes. The probes were hybridized to Atlas cDNA expression arrays containing 588 genes (Clontech) according to the manufacturer's suggestions.

Δ FosB-inducible transgenic mice were either fed 100 mg l⁻¹ doxycycline or denied doxycycline for 12 weeks. Adult male Sprague-Dawley rats initially weighing between 160 and 240 g were injected with cocaine or vehicle (intraperitoneal, i.p.; 20 mg kg⁻¹, 0.9% NaCl) at the same time each day for 8 d. All analyses were initiated 12 h after the final dose. For *in situ* hybridization studies, [α -³⁵S]UTP-labelled riboprobes were prepared by *in vitro* transcription from cDNA clones corresponding to full-length clones of rat Cdk5 and p35²². Cryostat sections were prepared and hybridized as previously described²³. After hybridization, the sections were exposed to Biomax MR film (Kodak) for 2–6 d and analysed with a Microcomputer Imaging Device system (M4, Imaging Research, Inc.). Statistical analyses of the data were performed using the two-tailed unpaired Student's *t*-test. Striatal tissue was rapidly dissected on ice and homogenized in boiling lysis buffer containing 1% SDS and 50 mM sodium fluoride with sonication. Striatal slice preparation and treatment were conducted as previously described⁷. Use of antibodies to phospho-Thr 34 DARPP-32⁸, phospho-Thr 75 DARPP-32⁷, phospho-Ser 55 ARPP-21¹⁰, phospho-Ser 88 ARPP-16^{7,9} and phospho-Ser 845 GluR1²⁴ for immunoblotting has been previously described.

Immunocytochemistry was performed as previously described²⁵ except that 0.3% glutaraldehyde was added to the fixative solution. Vibratome sections were incubated with a mixture of antibodies against DARPP-32 and Cdk5 (C-8, Santa Cruz Biotechnologies, Inc.) overnight at 25 °C. After a 30-min rinse in buffer, sections were incubated in a mixture of fluorescein-isothiocyanate-labelled goat antimouse IgG and Cy3-labelled goat anti-rabbit IgG for 1 h, followed by three 30-min washes in buffer. Sections were mounted in Vectashield and analysed with a Zeiss 410 confocal laser scanning microscope. Images were acquired separately in each channel (dual scan mode) to eliminate the possibility of signal bleed-over from one channel to the other.

Drug-induced alterations in cocaine sensitization were measured in adult male Sprague-Dawley rats according to published procedures for surgery, drug infusion, apparatus and behavioural methods^{26,27}. Coordinates for the nucleus accumbens were anterior-posterior 1.7 from bregma, medial-lateral $\pm$ 1.5 from midline, dorso-ventral –6.0 from skull. Cocaine hydrochloride was administered at 15 mg kg⁻¹ i.p. in sterile 0.9% sodium chloride. Roscovitine, olomoucine and iso-olomoucine were dissolved in sterile phosphate-buffered saline:DMSO (50:50) and 40 nmol were microinfused in 0.5 μ l over a 2-min period (compare ref. 28). One week after surgery and 2 d after habituation to the locomotor chambers, subjects received five daily injections of cocaine (15 mg kg⁻¹, i.p.) or saline 20 min after bilateral intra-accumbens vehicle, roscovitine, olomoucine or iso-olomoucine infusions. Drugs were given at the same time each day. Subjects were then placed into the chambers and activity was monitored for 60 min.

For electrophysiological studies, acutely dissociated striatal neurons were prepared as previously described²⁹. Whole-cell recordings of voltage-gated AMPA/kainate-activated current (100 μ M kainate) were measured using standard whole-cell voltage-clamp techniques³⁰. Summary data are presented in box plot format.

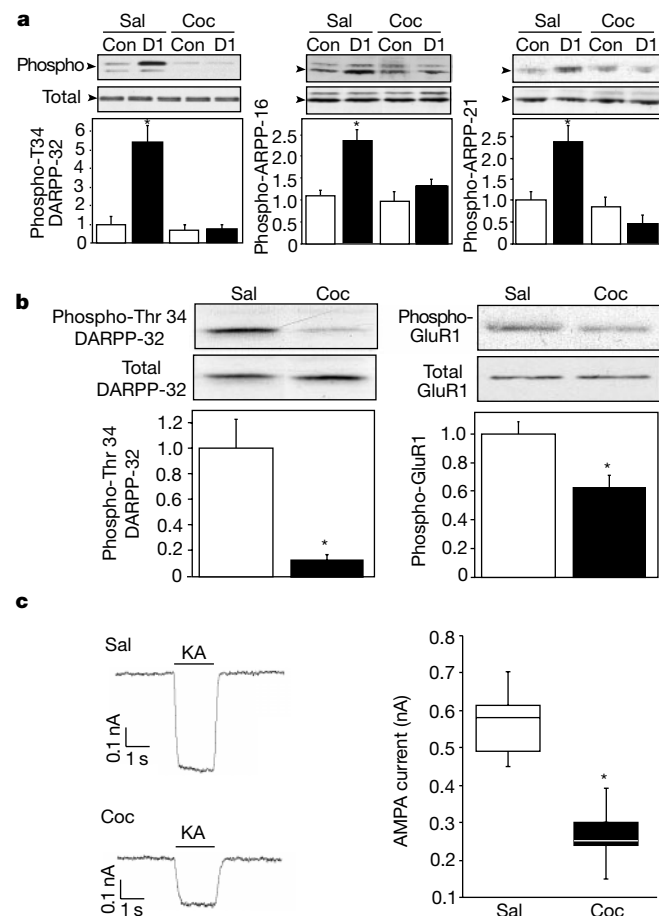


Figure 4 Effects of chronic cocaine exposure on dopamine/PKA signalling in the striatum. **a**, Effect of the selective D1 agonist SKF 81297 (D1), in comparison to untreated controls (Con), on PKA phosphorylation of DARPP-32 (left), ARPP-21 (middle) and ARPP-16 (right) in striatal slices from saline- and cocaine-treated rats. Representative blots are shown for proteins phosphorylated by PKA (top panels) and total proteins (middle panels). Levels are shown in the bottom panels. **b**, Effects of chronic cocaine exposure on PKA phosphorylation of DARPP-32 and GluR1 in tissue *in vivo*. Data in **a** and **b** represent means $\pm$ s.e.m., asterisk indicates $P < 0.01$ versus controls, Mann-Whitney non-parametric *t*-test, $n = 6$. **c**, Effect of chronic cocaine on ligand-gated AMPA/kainate current. Kainate (KA)-sensitive AMPA current recordings in saline- or cocaine-treated rats (left) and statistical analyses (right) are shown. Data represent means $\pm$ s.e.m., asterisk indicates $P < 0.01$ versus controls, unpaired *t*-test, $n = 6$.

Received 6 November; accepted 21 December 2000.

- Berke, J. D. & Hyman, S. E. Addiction, dopamine, and the molecular mechanisms of memory. *Neuron* **25**, 515–532 (2000).
- Nestler, E. J. Molecular mechanisms of opiate and cocaine addiction. *Curr. Opin. Neurobiol.* **7**, 713–719 (1997).
- Hope, B. T. *et al.* Induction of a long-lasting AP-1 complex composed of altered Fos-like proteins in brain by chronic cocaine and other chronic treatments. *Neuron* **13**, 1235–1244 (1994).
- Kelz, M. B. *et al.* Expression of the transcriptional factor Δ FosB in the brain controls sensitivity to cocaine. *Nature* **401**, 272–276 (1999).
- Hiroi, N. *et al.* FosB mutant mice: Loss of chronic cocaine induction of Fos-related proteins and heightened sensitivity to cocaine's psychomotor and rewarding effects. *Proc. Natl Acad. Sci. USA* **94**, 10397–10402 (1997).
- Chen, J. *et al.* Transgenic animals with inducible, targeted gene expression in the brain. *Mol. Pharmacol.* **54**, 495–503 (1998).
- Bibb, J. A. *et al.* Phosphorylation of DARPP-32 by Cdk5 modulates dopamine signalling in neurons. *Nature* **402**, 669–671 (1999).
- Snyder, G. L. *et al.* Phosphorylation of DARPP-32 and protein phosphatase inhibitor-1 in rat choroid plexus: Regulation by factors other than dopamine. *J. Neurosci.* **12**, 3071–3083 (1992).
- Dulubova, I. *et al.* ARPP-16/ARPP-19: a highly conserved family of cAMP-regulated phosphoproteins. *J. Neurochem.* (in the press).
- Caporaso, G. *et al.* Dopamine and drugs of abuse modulate phosphorylation of ARPP-21, a cyclic AMP-regulated phosphoprotein enriched in the neostriatum. *Neuropharmacology* **39**, 1637–1644 (2000).
- Roche, K. W., O'Brien, R. J., Mammen, A. L., Bernhardt, J. & Haganir, R. L. Characterization of multiple phosphorylation sites on the AMPA receptor GluR1 subunit. *Neuron* **16**, 1179–1188 (1996).
- White, F. J., Hu, X. -T., Zhang, X. -F. & Wolf, M. E. Repeated administration of cocaine or amphetamine alters neuronal responses to glutamate in the mesoaccumbens dopamine system. *J. Pharmacol. Exp. Ther.* **273**, 445–454 (1995).
- Zhang, X. -F., Hu, X. -T. & White, F. J. Whole-cell plasticity in cocaine withdrawal: Reduced sodium currents in nucleus accumbens neurons. *J. Neurosci.* **18**, 488–498 (1998).
- Dreher, J. K. & Jackson, D. M. Role of D1 and D2 dopamine receptors in mediating locomotor activity elicited from the nucleus accumbens of rats. *Brain Res.* **487**, 267–277 (1989).
- Self, D. W., Barnhart, W. J., Lehman, D. A. & Nestler, E. J. Opposite modulation of cocaine-seeking behavior by D1- and D2-like dopamine receptor agonists. *Science* **271**, 1586–1589 (1996).
- Xu, M. *et al.* Elimination of cocaine-induced hyperactivity and dopamine-mediated neurophysiological effects in dopamine D1 receptor mutant mice. *Cell* **79**, 945–955 (1994).
- Nishi, A. *et al.* Amplification of dopaminergic signalling by a novel positive feedback loop. *Proc. Natl Acad. Sci. USA* **97**, 12840–12845 (2000).
- Hiroi, N. *et al.* Neuronal and behavioural abnormalities in striatal function in DARPP-32-mutant mice. *Eur. J. Neurosci.* **11**, 1114–1118 (1999).
- Chae, T. *et al.* Mice lacking p35, a neuronal specific activator of cdk5, displays cortical lamination defects, seizures, and adult lethality. *Neuron* **18**, 29–42 (1997).
- Ohshima, T. *et al.* Targeted disruption of the cyclin-dependent kinase 5 gene results in abnormal corticogenesis, neuronal pathology and perinatal death. *Proc. Natl Acad. Sci. USA* **93**, 11173–11178 (1996).
- Patrick, G. N. *et al.* Conversion of p35 to p25 deregulates Cdk5 activity and promotes neurodegeneration. *Nature* **402**, 615–622 (1999).
- Zheng, M., Leung, C. L. & Liem, R. K. Region-specific expression of cyclin-dependent kinase 5 (cdk5) and its activators, p35 and p39, in the developing and adult rat central nervous system. *J. Neurobiol.* **35**, 141–159 (1998).
- Le Moine, C. & Bloch, B. D1 and D2 dopamine receptor gene expression in the rat striatum: sensitive cRNA probes demonstrate prominent segregation of D1 and D2 mRNAs in distinct neuronal populations of the dorsal and ventral striatum. *J. Comp. Neurol.* **355**, 418–426 (1995).
- Lee, H. -K., Kameyama, K., Haganir, R. L. & Bear, M. F. NMDA induces long-term synaptic depression and dephosphorylation of the GluR1 subunit of AMPA receptors in hippocampus. *Neuron* **21**, 1151–1162 (1998).
- Ouimet, C. C. Immunocytochemical localization of protein phosphatases and their inhibitors. *Neuroprotocols* **6**, 84–90 (1995).
- Horger, B. A. *et al.* Enhancement of locomotor activity and conditioned reward to cocaine by brain-derived neurotrophic factor. *J. Neurosci.* **19**, 4110–4122 (1999).
- Taylor, J. R. & Horger, B. A. Enhanced responding for conditioned reward produced by intra-accumbens amphetamine is potentiated after cocaine sensitization. *Psychopharmacology* **142**, 31–40 (1999).
- Punch, L., Self, D., Nestler, E. J. & Taylor, J. R. Opposite modulation of opiate withdrawal behaviors on microinfusion of a protein kinase A inhibitor versus activator into the locus coeruleus or periaqueductal gray. *J. Neurosci.* **17**, 8520–8527 (1997).
- Surmeier, D. J., Vargas, J., Hemmings, H. C. Jr, Nairn, A. C. & Greengard, P. Modulation of calcium currents by a D1 dopaminergic protein kinase/phosphatase cascade in rat neostriatal neurons. *Neuron* **14**, 385–397 (1995).
- Hamill, O. P., Marty, A., Neher, E., Sakmann, B. & Sigworth, F. J. Improved patch-clamp techniques for high resolution current recording from cells and cell-free membrane patches. *Pflügers Arch.* **391**, 85–100 (1981).

Acknowledgements

We thank R. Liem for Cdk5 and p35 cDNA probes; L. Meijer for roscovitine and olomoucine; and R. L. Haganir for phospho-Ser845 GluR1 antibody. We also thank V. Phantharagasy, V. Stewart and E. Griggs for technical assistance. This work was supported by a NARSAD young investigator award (to J.A.B.), a postdoctoral fellowship from Stiftelsen för Internationalisering av högre utbildning och forskning (to P.S.), and grants from the USPHS (J.R.T., E.J.N., G.L.S., A.C.N. and P.G.).

Correspondence and requests for materials should be addressed to J.A.B. (e-mail: bibbj@rockvax.rockefeller.edu).

BRI1 is a critical component of a plasma-membrane receptor for plant steroids

Zhi-Yong Wang*, Hideharu Seto†, Shozo Fujioka†, Shigeo Yoshida† & Joanne Chory*

*Howard Hughes Medical Institute and Plant Biology Laboratory, The Salk Institute for Biological Studies, 10010 N. Torrey Pines Road, La Jolla, California 92037, USA

† Plant Functions Lab, RIKEN (The Institute of Physical and Chemical Research), Wako-shi, Saitama 351-0198, Japan

Most multicellular organisms use steroids as signalling molecules for physiological and developmental regulation. Two different modes of steroid action have been described in animal systems: the well-studied gene regulation response mediated by nuclear receptors^{1,2}, and the rapid non-genomic responses mediated by proposed membrane-bound receptors^{3,4}. Plant genomes do not seem to encode members of the nuclear receptor superfamily⁵. However, a transmembrane receptor kinase, brassinosteroid-insensitive1 (BRI1), has been implicated in brassinosteroid responses^{6,7}. Here we show that BRI1 functions as a receptor of brassinolide, the most active brassinosteroid. The number of brassinolide-binding sites and the degree of response to brassinolide depend on the level of BRI1 protein. The brassinolide-binding activity co-immunoprecipitates with BRI1, and requires a functional BRI1 extracellular domain. Moreover, treatment of *Arabidopsis* seedlings with brassinolide induces autophosphorylation of BRI1, which, together with our binding studies, shows that BRI1 is a receptor kinase that transduces steroid signals across the plasma membrane.

Brassinosteroids (BRs) are involved in a wide range of plant developmental processes⁸. Mutant plants deficient in BR biosynthesis, such as *det2*, show phenotypes of dwarfism, delayed senescence, reduced fertility, and light-independent development^{9,10}. Mutations in the *BRI1* gene cause BR-insensitivity and morphological phenotypes nearly identical to BR biosynthetic mutants, suggesting an important and specific role for *BRI1* in BR perception or signal transduction^{11,12}. The *BRI1* gene encodes a receptor kinase that has an extracellular domain containing 25 leucine-rich repeats (LRRs), which are interrupted by a 70-amino-acid island, a transmembrane domain, and a cytoplasmic kinase domain with serine/threonine specificity^{6,13}. The structure of BRI1 and its plasma membrane localization¹³ support the hypothesis that BRI1 interacts with an extracellular ligand, which is either BR itself or a secondary signal generated by BR perception, and the signal is transduced through the kinase. Consistent with this hypothesis, recent studies using a chimaeric receptor approach showed that the extracellular domain of BRI1 could confer brassinolide (BL) responsiveness to the intracellular kinase domain of a heterologous LRR kinase⁷.

To test whether BL is the ligand that directly activates the BRI1 receptor kinase, we first analysed the effect of overexpression of BRI1 on BL binding activity in membrane fractions. Transgenic *Arabidopsis* plants overexpressing a BRI1–GFP fusion protein¹³ (GFP, green fluorescent protein) showed reduced inhibition of hypocotyl growth by a BR biosynthesis inhibitor¹⁴ (Fig. 1a, b). They also had longer petioles, similar to plants overexpressing the BR biosynthetic enzyme DWF4¹⁵ (Z.W. and J.C., unpublished data) (Fig. 1c). These phenotypes are consistent with the interpretation that overexpression of the BRI1–GFP protein increases the response of *Arabidopsis* to BRs. We observed a dramatic increase of BL binding activity in the membrane fractions of the BRI1–GFP transgenic plants (Fig. 1d). The increase of binding was due to an

increase of binding sites (maximum number of binding sites, $B_{\max} = 2.66$ pmol per mg membrane protein compared to 0.23 pmol per mg membrane protein), with similar binding affinities (dissociation constant $K_d = 7.4 \pm 0.9$ nM compared to 10.8 ± 3.2 nM) (Fig. 1e). Such K_d values are consistent with physiological concentrations of BL^{16,17} and coincide with the BL concentration that induces 50% of the maximum growth response in BL-deficient mutants⁹.

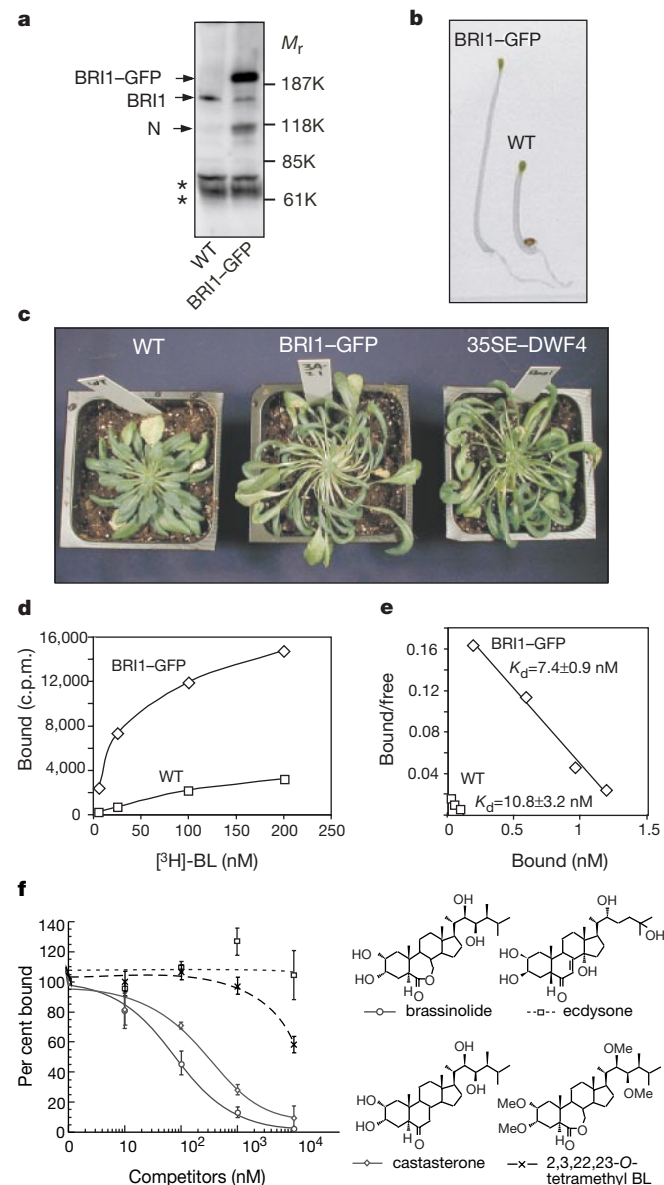


Figure 1 Overexpression of BRI1-GFP increases cell elongation and increases the number of brassinolide (BL) binding sites in membrane fractions. **a**, Proteins from wild-type and BRI1-GFP transgenic plants probed with anti-BRI1N antibody after western blotting. N, a cleaved fragment of BRI1's extracellular domain. Asterisks, non-specific bands. **b**, Wild-type (WT) and BRI1-GFP plants grown on 2 μM brassinazole in the dark for 6 d. **c**, A wild-type plant, a BRI1-GFP transgenic plant, and a mutant plant overexpressing the *DWF4* gene (35SE-DWF4) grown for 45 d in cycles of 9 h light/15 h dark. **d**, Specific [³H]-BL binding to microsomal fractions of wild-type (WT) and BRI1-GFP plants was determined by subtracting the binding in the presence of 100-fold unlabelled BL from the total binding in the absence of cold competitor. Representative data of one of three repeat experiments are shown. **e**, Scatchard plot of the binding data in **d**. The K_d values were calculated from data of three experiments, with correlation coefficient $R^2 = 0.998$ for BRI1-GFP and 0.983 for wild-type samples. **f**, Competition for [³H]-BL binding to membrane fractions of BRI1-GFP plants by brassinolide, castasterone, ecdysone and 2,3,22,23-O-tetramethylbrassinolide. Structures of the competitors are shown.

We determined the specificity of the BL binding activity by comparing the relative binding affinity for several steroid compounds in binding competition assays (Fig. 1f). Binding of [³H]-BL to the membrane fraction of BRI1-GFP plants was effectively competed by unlabelled BL (50% inhibition concentration, IC_{50} , 80 nM), less effectively by castasterone (IC_{50} , 340 nM), and not competed by 2,3,22,23-O-tetramethylbrassinolide (Me-BL; $IC_{50} > 10$ μM) and ecdysone ($IC_{50} > 10$ μM). The relative binding affinity (the ratio between IC_{50} of a competitor and that of BL) of castasterone is about 4–5 times lower than BL, and this is consistent with castasterone being about 5 times less active than BL in bioassays¹⁸. The lack of competition by Me-BL and ecdysone is consistent with their lack of biological activity in

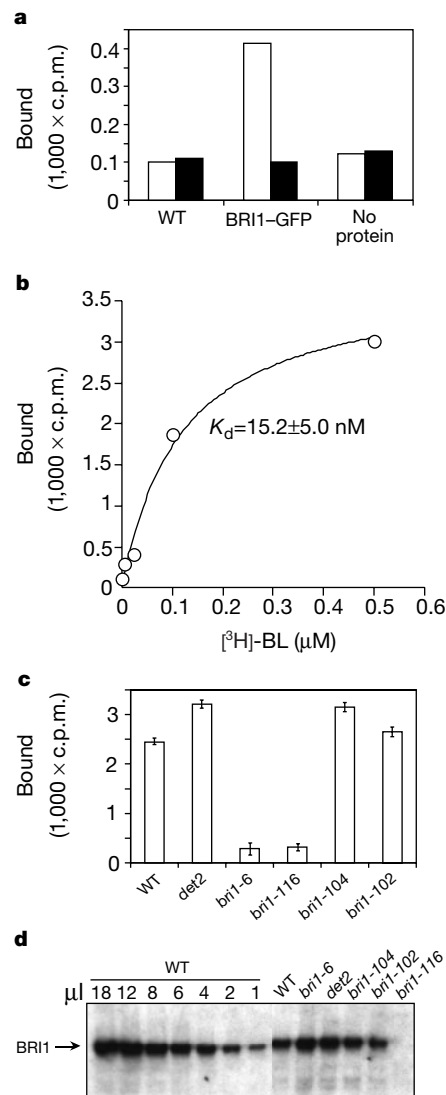


Figure 2 BRI1 binds to BL. **a**, Proteins immunoprecipitated with anti-GFP antibodies from extracts of wild-type or BRI1-GFP plants were assayed for [³H]-BL binding activity in the absence (open bars) or presence (filled bars) of 5 μM unlabelled BL. No protein, binding assays with protein A beads as control. **b**, Specific and saturation [³H]-BL binding to immunoprecipitated BRI1-GFP protein. **c**, Mutations in the extracellular domain of BRI1, but not the kinase domain, reduce BL binding. Specific [³H]-BL binding to microsomal fractions of wild type (WT), *det2*, *bri1-6*, *bri1-116*, *bri1-104* and *bri1-102* mutant plants. Data were normalized to the relative BRI1 protein levels determined by quantitative western blotting (**d**), except for *bri1-116*. **d**, Protein immunoblot showing the BRI1 protein levels in the membrane fractions used in the binding assays of **c**. Varying loading of the wild-type sample was used to generate a standard curve, which was used to determine the relative level of BRI1 in the *bri1* mutant samples (6 μl per lane).

plants¹⁹ (S.F., unpublished results). Such a specificity and high affinity for biologically active BRs indicate that this BL binding activity accounts for the BL-induced biological responses.

A specific BL binding activity was detected after the BRI1–GFP proteins were immunoprecipitated using anti-GFP antibodies (Fig. 2a, b). No specific BL binding activity was immunoprecipitated from wild-type *Arabidopsis* plants using the same antibodies (Fig. 2a), indicating that the BL binding activity is specific to the BRI1–GFP protein. The immunoprecipitated binding activity has a similar dissociation constant ($K_d = 15.2 \pm 5$ nM) as determined for membrane fractions (Fig. 2b). These results show that BRI1 either binds BL directly or is a limiting component of a receptor complex for BL in plant cells.

Mutations in large numbers of *bri1* alleles implicate the functional importance of the cytoplasmic kinase domain and the 70-amino-acid island of BRI1's extracellular domain^{6,13,17}. We found that *bri1* mutants with missense mutations in the kinase domain (*bri1-104*, Ala1031Thr) or in a region of the extracellular domain near the transmembrane domain (*bri1-102*, Thr750Ile) have BL binding activities similar to wild type and the biosynthetic mutant *det2* (Fig. 2c, d). In contrast, a missense mutation (*bri1-6*, Gly644Asp) and a nonsense mutation that results in a truncated protein at position Gln 583 (*bri1-116*), both in the 70-amino-acid island region, greatly reduced the BL binding activity (Fig. 2c, d). These results provide direct evidence that the 70-amino-acid island region of BRI1's extracellular domain is required for BL binding to the receptor on the cell membrane.

We tested whether BL-binding activates BRI1's kinase. Receptor activation often involves auto-phosphorylation, which can lead to a

change of mobility in SDS–polyacrylamide gel electrophoresis (PAGE). *Arabidopsis* seedlings grown in the presence of the BR biosynthetic inhibitor brassinazole were treated with BL and analysed by immunoblotting (Fig. 3). Treatment of wild-type seedlings with 1 μ M BL for 1 h caused a shift of BRI1 from a faster to a slower migrating band, compared with an untreated sample or a sample treated with mock solution (Fig. 3a). Phosphatase treatment of the BL-treated samples shifted the slower band back to the fast migrating band, suggesting that the shift of mobility represents BRI1 phosphorylation (Fig. 3b). We also observed such a BL-induced BRI1 mobility shift in the BL biosynthetic mutant *det2*, but not in the *bri1-117* mutant (Fig. 3c), which contains a mutation that abolishes BRI1's *in vitro* kinase activity (data not shown). Our data indicate that BL induction of BRI1 phosphorylation requires the kinase activity of BRI1, suggesting that BL-binding induces autophosphorylation of BRI1.

Our identification of the receptor kinase BRI1 as a plant steroid receptor illustrates the function of a member of the largest family of receptor kinases in *Arabidopsis*. The *Arabidopsis* genome sequence revealed 174 LRR-receptor kinases⁵, of which only a few are known for their biological functions^{20–22}, and only one, CLV1, has been characterized at the biochemical level²⁰. The mechanism by which LRR kinase is activated by ligand binding may be shared by other LRR-receptor kinases, as suggested by the BR activation of a BRI1–Xa21 chimaeric receptor⁷. However, BRI1 seems to differ from CLV1, which has recently been shown to require its own kinase activity for binding to its peptide ligand²⁰.

Our results also reveal a new mechanism of steroid signalling. Steroid hormones are generally known to pass freely across plasma membranes into animal cells, where they bind to members of the nuclear receptor superfamily of ligand-dependent transcription factors^{1,2}. In contrast, the *Arabidopsis* genome does not seem to encode members of this family of proteins⁵. The near identical phenotypes of *bri1* to BR-biosynthetic mutants and our results presented here indicate that plants perceive steroids at the cell surface and that BRI1 is likely to be the primary BR receptor in *Arabidopsis*. A similar BL signalling mechanism is apparently conserved in other plants, as a homologue of BRI1 was shown recently to be required for BR responses in rice²³. Such a cell-surface signalling mechanism may not be unique to plant steroids. In fact, membrane-initiated steroid responses have been observed in many animal systems, and signalling molecules such as calcium, inositol phosphates, cyclic AMP, G proteins and various kinases have been implicated³. However, little is known about the membrane-bound steroid receptors that initiate these signalling cascades in animal cells⁴. Considering the conservation of the steroid biosynthetic enzymes²⁴ and similar roles of steroids in growth and differentiation in plants and animals, it will be interesting to see if a similar membrane-based steroid signalling mechanism exists in both the animal and plant kingdoms. □

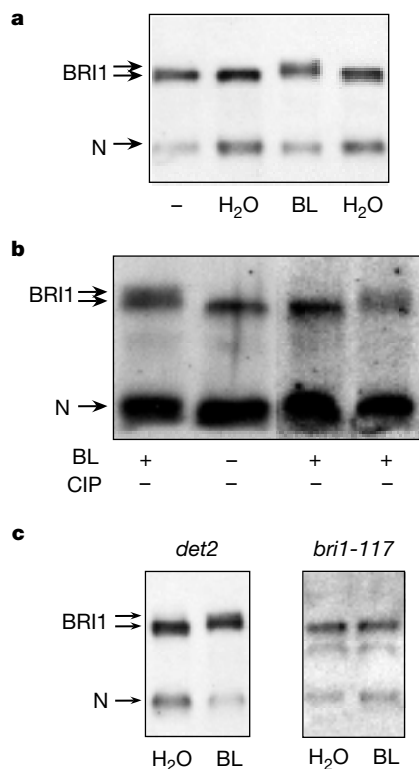


Figure 3 BL induces phosphorylation of BRI1 in plants. **a**, Five-day-old wild-type seedlings grown in the dark on medium containing 1 μ M brassinazole were untreated (–), treated with 1 μ M BL in water (BL), or with water only (H₂O) for 1 h. Proteins were analysed by 4% SDS–PAGE, blotted, and probed with anti-BRI1N antibody. **b**, Proteins of the BL-treated wild-type sample were treated with alkaline phosphatase (CIP), and analysed by western blotting as in **a**. **c**, A mutation in the kinase domain abolishes BL-activation of BRI1 phosphorylation. The *det2* and *bri1-117* mutant seedlings were treated and analysed as in **a**. N, cleavage product of BRI1's extracellular domain.

Methods

BL-binding assays

Tritium-labelled BL was made by American RadioChemicals using tritium reduction of 25,26-dehydrobrassinolide²⁵. The specific activity of [³H]-BL was estimated to be 50 Ci mmol^{–1}, and the correct structure was confirmed by tritium NMR analysis. Biological activity of the [³H]-BL was determined by rescue of the *det2* mutant (data not shown). Plant microsomal fractions were prepared from *Arabidopsis* seedlings grown in a cycle of 9 h light/15 h dark for about 6 weeks, following a protocol described previously⁷. Membrane pellets were resuspended at a protein concentration of 2 mg ml^{–1} in BL-binding buffer (0.25 M mannitol, 10 mM Tris-2-[N-morpholino] ethanesulphonic acid (MES), pH 5.7, 5 mM MgCl₂, 0.1 mM CaCl₂, and protease inhibitor cocktail (Sigma). Each BL binding assay contains 50 μ l membrane suspensions, 50 nM or indicated amount of [³H]-BL, without or with 100-fold excess unlabelled BL or indicated amount of unlabelled steroids, 1 mg ml^{–1} BSA and BL-binding buffer in 100 μ l total volume. The binding reactions were incubated for 1 h or indicated time at 25 °C. The bound and free [³H]-BL were separated by filtering the mixture through a glass-fibre filter (Whatman, GF/F) and washing with 10 ml ice-cold BL-binding buffer, and were quantified by scintillation counting. Binding kinetic studies showed that the specific BL binding reaches equilibrium

within 15 min of incubation; the binding is highly reversible and 50% competition by unlabelled BL was achieved in less than 1 min (data not shown). For binding assays with immunoprecipitated proteins, proteins were extracted with BRI1 extraction buffer (2 ml per g tissue) (50 mM Tris-HCl, pH 7.5, 10 mM NaCl, 5 mM EDTA, 1% Triton X-100, and protease inhibitor cocktail), and GFP-tagged proteins were immunoprecipitated using anti-GFP antibodies (Molecular Probes, 1 µl per ml extract) and protein A agarose beads (10 µl per ml extract, Pierce). BL binding assays with immunoprecipitant/agarose beads were under the same conditions as for membrane fractions. Specific binding was determined by subtracting the binding in the presence of 100 fold unlabelled BL from total binding. Binding data were analysed and plotted using KaleidaGraph software (Synergy Software).

Immunoblotting and assays for BL-induced BRI1 phosphorylation

A peptide containing the first 106 amino acids of BRI1 (excluding the signal peptide) was expressed in *Escherichia coli* and purified as a maltose binding protein (MBP) fusion. This MBP-BRI1N fusion protein was used as an antigen for generating the anti-BRI1N antibodies in rabbits, and for affinity purification of the antibodies. The identities of the BRI1-containing bands detected by the anti-BRI1N antibodies on immunoblots were determined by comparing wild type, BRI1-GFP and *bri1-116* (a nonsense mutant) samples. To test BL induction of BRI1 phosphorylation, *Arabidopsis* seedlings were grown on MS medium plates containing 1 µM brassinazole in the dark for 4 d, submerged in water or in 1 µM BL solution for 1 min, then put back on MS plates without or with 1 µM BL for 1 h. Samples were analysed by SDS-PAGE using 4% gels and western blotting as described above⁷. For phosphatase treatment, SDS was removed from protein samples using the SDS-OUT kit (Pierce), and the proteins were then treated with alkaline phosphatase (Boehringer Mannheim) under conditions recommended by the manufacturer and described elsewhere²⁶.

Received 3 October; accepted 20 December 2000.

- Beato, M., Herrlich, P. & Schutz, G. Steroid hormone receptors: many actors in search of a plot. *Cell* **83**, 851–857 (1995).
- Mangelsdorf, D. J. *et al.* The nuclear receptor superfamily: The second decade. *Cell* **83**, 835–839 (1995).
- Wehling, M. Specific, nongenomic actions of steroid hormones. *Annu. Rev. Physiol.* **59**, 365–393 (1997).
- Schmidt, B. M. *et al.* Rapid, nongenomic steroid actions: A new age? *Front. Neuroendocrinol.* **21**, 57–94 (2000).
- The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
- Li, J. & Chory, J. A putative leucine-rich repeat receptor kinase involved in brassinosteroid signal transduction. *Cell* **90**, 929–938 (1997).
- He, Z. *et al.* Perception of brassinosteroids by the extracellular domain of the receptor kinase BRI1. *Science* **288**, 2360–2363 (2000).
- Mandava, N. B. Plant growth-promoting brassinosteroids. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **39**, 23–52 (1988).
- Li, J., Nagpal, P., Vitart, V., McMorris, T. C. & Chory, J. A role for brassinosteroids in light-dependent development of *Arabidopsis*. *Science* **272**, 398–401 (1996).
- Wang, Z.-Y. & Chory, J. In *Recent Advances in Phytochemistry* Vol. 34, *Evolution of Metabolic Pathways* (eds Romeo, J. T., Ibrahim, R., Varin, L. & DeLuca, V.) 409–431 (Elsevier Science, Oxford, 2000).
- Clouse, S. D., Langford, M. & McMorris, T. C. A brassinosteroid-insensitive mutant in *Arabidopsis thaliana* exhibits multiple defects in growth and development. *Plant Physiol.* **111**, 671–678 (1996).
- Schumacher, K. & Chory, J. Brassinosteroid signal transduction: still casting the actors. *Curr. Opin. Plant Biol.* **3**, 79–84 (2000).
- Friedrichsen, D. M., Joazeiro, C. A., Li, J., Hunter, T. & Chory, J. Brassinosteroid-insensitive-1 is a ubiquitously expressed leucine-rich repeat receptor Serine/Threonine kinase. *Plant Physiol.* **123**, 1247–1256 (2000).
- Asami, T. *et al.* Characterization of brassinazole, a triazole-type brassinosteroid biosynthesis inhibitor. *Plant Physiol.* **123**, 93–100 (2000).
- Choe, S. *et al.* The DWF4 gene of *Arabidopsis* encodes a cytochrome P450 that mediates multiple 22α-hydroxylation steps in brassinosteroid biosynthesis. *Plant Cell* **10**, 231–243 (1998).
- Clouse, S. & Sasse, J. Brassinosteroids: Essential regulators of plant growth and development. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **49**, 427–451 (1998).
- Noguchi, T. *et al.* Brassinosteroid-insensitive dwarf mutants of *Arabidopsis* accumulate brassinosteroids. *Plant Physiol.* **121**, 743–752 (1999).
- Fujioka, S., Noguchi, T., Takatsutod, S. & Yoshida, S. Activity of brassinosteroids in the dwarf rice lamina inclination bioassay. *Phytochemistry* **49**, 1841–1848 (1998).
- Luo, W., Janzen, L., Pharis, R. P. & Back, T. G. Bioactivity of brassinolide methyl ethers. *Phytochemistry* **49**, 637–642 (1998).
- Trotochaud, A. E., Jeong, S. & Clark, S. E. CLAVATA3, a multimeric ligand for the CLAVATA1 receptor-kinase. *Science* **289**, 613–617 (2000).
- Torii, K. U. *et al.* The *Arabidopsis* ERECTA gene encodes a putative receptor protein kinase with extracellular leucine-rich repeats. *Plant Cell* **8**, 735–746 (1996).
- Jinn, T. L., Stone, J. M. & Walker, J. C. HAESA, an *Arabidopsis* leucine-rich repeat receptor kinase, controls floral organ abscission. *Genes Dev.* **14**, 108–117 (2000).
- Yamamoto, C. *et al.* Loss of function of a rice brassinosteroid insensitive1 homolog prevents internode elongation and bending of the lamina joint. *Plant Cell* **12**, 1591–1605 (2000).
- Li, J., Biswas, M. G., Chao, A., Russell, D. W. & Chory, J. Conservation of function between mammalian and plant steroid 5α-reductases. *Proc. Natl Acad. Sci. USA* **94**, 3554–3559 (1997).
- Seto, H. *et al.* A general approach to synthesis of labeled brassinosteroids: preparation of [25,26,27-³H₃]brassinolide with 60% isotopic purity from the parent brassinolide. *Tetrahed. Lett.* **39**, 7525–7528 (1998).

26. Fankhauser, C. *et al.* PKS1, a substrate phosphorylated by phytochrome that modulates light signaling in *Arabidopsis*. *Science* **284**, 1539–1541 (1999).

Acknowledgements

We thank M. Chen for comments and L. Barden for technical assistance on the manuscript; D. Vafeados for technical assistance; and D. Friedrichsen for providing the BRI1-GFP line. This work was supported by a grant from the USDA and the Howard Hughes Medical Institute to J.C., and by a Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sports, and Culture of Japan to S.F. Z.W. is an NSF postdoctoral fellow and J.C. is an Associate Investigator of the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to J.C. (e-mail: chory@salk.edu).

Regulation of CD40 and CD40 ligand by the AT-hook transcription factor AKNA

Aisha Siddiqua*†, Jennifer C. Sims-Mourtada*†, Liliana Guzman-Rojas*†, Roberto Rangel*†, Christiane Guret‡, Vicente Madrid-Marina§, Yan Sun† & Hector Martinez-Valdez†

† Department of Immunology, Box 178, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, USA

‡ Laboratory for Immunology Research, Schering-Plough, 27 Chemin des Peupliers, 69571 Dardilly, Cedex, France

§ Centro de Investigacion Sobre Enfermedades Infecciosas, Instituto Nacional de Salud Publica, Ave. Universidad 655, Cuernavaca, Morelos, 62508, Mexico

* These authors contributed equally to this work.

Proteins containing AT hooks bind A/T-rich DNA through a nine-amino-acid motif and are thought to co-regulate transcription by modifying the architecture of DNA, thereby enhancing the accessibility of promoters to transcription factors^{1,2}. Here we describe AKNA, a human AT-hook protein that directly binds the A/T-rich regulatory elements of the promoters of CD40 and CD40 ligand (CD40L) and coordinately regulates their expression. Consistent with its function, AKNA is a nuclear protein that contains multiple PEST protein-cleavage motifs, which are common in regulatory proteins with high turnover rates³. AKNA is mainly expressed by B and T lymphocytes, natural killer cells and dendritic cells. During B-lymphocyte differentiation, AKNA is mainly expressed by germinal centre B lymphocytes, a stage in which receptor and ligand interactions are crucial for B-lymphocyte maturation^{4–12}. Our findings show that an AT-hook molecule can coordinately regulate the expression of a key receptor and its ligand, and point towards a molecular mechanism that explains homotypic cell interactions.

Earlier studies of lymphocyte maturation indicated the possible existence of stage-specific molecules that could regulate ligand and receptor gene expression^{12–15}. To test this hypothesis, we performed a polymerase chain reaction with reverse transcriptase (RT-PCR) differential display¹⁶, using RNA from pure human B-lymphocyte subsets⁹. PCR products were purified, cloned, sequenced and compared against gene databases¹⁷, revealing that seven of the twelve PCR clones were known genes. AKNA and another protein, PRELI¹⁸, were differentially expressed by germinal centre B lymphocytes and were among the five new complementary DNAs.

Northern blotting revealed a 4-kilobase (kb) AKNA messenger RNA predominantly expressed by lymphoid tissues (Fig. 1). The expression was highest in the spleen, lymph nodes and peripheral

blood leukocytes (PBLs), lower in the thymus and undetectable in fetal liver and in adult bone marrow. These data indicate that, in the adult, AKNA expression may be relevant in secondary lymphoid organs. Consistent with its transcription pattern, AKNA was expressed by B and T lymphocytes, NKC and CD1a⁺CD14⁻ but not CD1a⁻CD14⁺ dendritic cells (Fig. 2). These findings (Figs 1, 2) were confirmed using a full-length cDNA probe.

We used the AKNA PCR probe to screen T- and B-lymphocyte cDNA libraries (Fig. 2) and identified, cloned and sequenced three overlapping cDNAs. Alignment of the three clones revealed a 3.7-kb sequence similar to AKNA's transcript. A translation initiation codon was followed by an 1.9-kb open reading frame (ORF) encoding a 633-amino-acid protein with a relative molecular mass of approximately 69,000 (Fig. 3). AKNA depicted a nine-amino-acid domain, RTRGRPADS, that satisfies the consensus requirements of an AT hook DNA-binding motif (Fig. 3). The AT hook is a small motif with a typical sequence signature, in which the tripeptide GRP is the centre of the DNA-binding domain². AT-hook-containing proteins have been identified in a variety of DNA-binding proteins¹⁹. AKNA also possesses three PEST regions³. Thus AKNA is probably an AT hook DNA-binding factor with a high turnover rate whose expression and function localizes in the nucleus. AKNA's subcellular localization was determined by transfection of the B-lymphoma cell line JY (ref. 20), using a green fluorescent protein (GFP) tag. Figure 4a, b shows that whereas the GFP control transfectants showed a homogeneous fluorescence distribution, AKNA-GFP was selectively nuclear. The co-localization of AKNA-GFP with the DNA-specific fluorochrome propidium iodide (Fig. 4c, d) further indicated that AKNA is a nuclear protein. These data indicate that AKNA may be involved in B-lymphocyte gene expression.

The predominant expression in secondary lymphoid organs (Fig. 1) indicated that AKNA could be important in antigen-dependent immune responses. Accordingly, immunohistochemical analysis showed that AKNA is predominantly expressed within the germinal centre microenvironment of secondary lymphoid organs (Fig. 5a), where antigen-dependent maturation of B lymphocytes takes place. To determine whether AKNA is expressed by B lymphocytes, we purified naive, germinal centre and memory cell subsets⁹ and showed that AKNA is predominantly expressed by germinal centre B lymphocytes (Fig. 5c), indicating

that AKNA may be important in the physiology of germinal centre reaction.

Within germinal centres, B lymphocytes are destined to die by apoptosis mediated by Fas signalling^{21,22} (Fig. 5d), unless they are positively selected by antigens and by signals initiated by CD40-CD40L interactions^{12,23,24}. In the absence of T cells during clonal selection at the basal light zone, concomitant expression of CD40 and CD40L by germinal centre B lymphocytes could be essential for the rescue of the antigen-selected repertoire. The capacity of germinal centre B lymphocytes to co-express the receptor and its ligand (Fig. 5d) agrees with this proposal. The active engagement of B lymphocytes in homotypic ligand-receptor interactions^{25,26} further supports the finding of the concomitant expression of CD40 and CD40L by germinal centre B lymphocytes. However, the mechanisms that coordinately regulate their transcription remain to be determined.

We used a commercial cDNA expression array to identify putative AKNA targets and observed a consistent upregulation of CD40 and downregulation of glutathione *S*-transferase²⁷ (GST) P1 expression

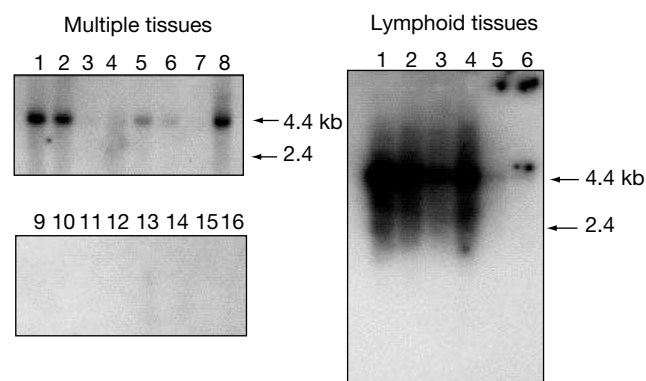


Figure 1 AKNA is predominantly expressed by secondary lymphoid organs. Northern blot analysis of AKNA expression in multiple tissues and by distinct lymphoid organs is indicated. Commercial northern blots were hybridized with a ³²P-labelled AKNA cDNA probe, which originated from a differentially expressed transcript by the germinal centre subset. Commercial blots are routinely normalized for equal housekeeping (β -actin or ubiquitin) gene expression. Multiple tissues: 1, spleen; 2, thymus; 3, prostate; 4, testis; 5, ovary; 6, intestine; 7, colon; 8, PBL; 9, heart; 10, brain; 11, placenta; 12, lung; 13, liver; 14, skeletal muscle; 15, kidney; 16, pancreas. Lymphoid tissues: 1, spleen; 2, lymph node; 3, thymus; 4, PBL; 5, bone marrow; 6, fetal liver.

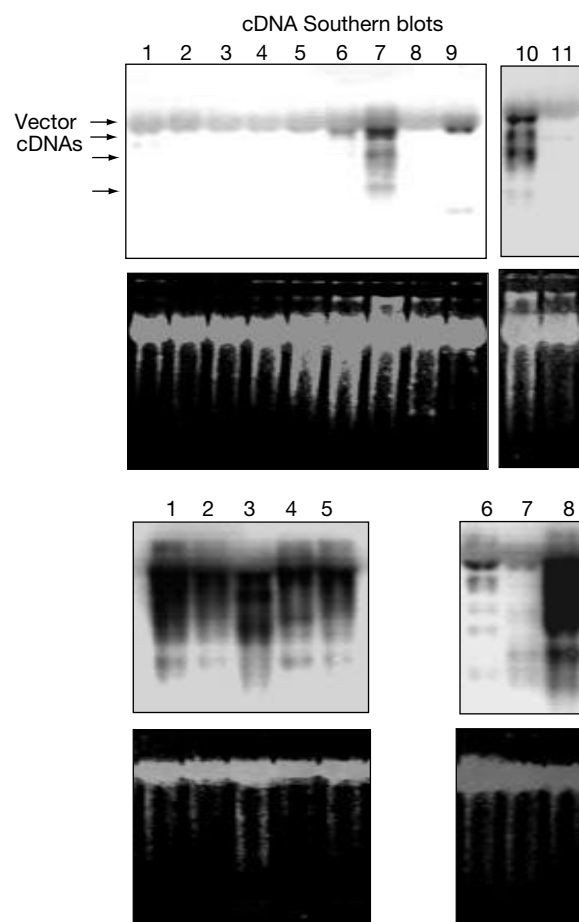


Figure 2 Expression of AKNA by B and T lymphocytes, NKC and CD1a⁺CD14⁻ cells. A cDNA library Southern blot analysis was carried out to analyse cell lineage-specific expression of AKNA and to identify a library source for the cloning of full-length AKNA cDNA. A ladder of positively hybridized cDNA fragments (including full lengths) is expected. The relative position of the vector (4.3 kb) is indicated. AKNA cDNA fragments are marked by arrows. Numbers identify the cell lineage tested and letters indicate multiple samplings. Ethidium bromide staining gels are controls for equal sample loading. Top: 1, gallbladder; 2, small intestine; 3, adipose tissue; 4, ovary; 5, uterus; 6, testis; 7, spleen; 8, placenta; 9, tonsil B cells; 10, CD1a⁺CD14⁻ dendritic cells; 11, CD1a⁻CD14⁺ dendritic cells. Bottom (lowercase letters indicate multiple samples): 1, TH1a; 2, TH1b; 3, TH1c; 4, TH2a; 5, TH2b; 6, NKCa; 7, NKCb; 8, NKCc.

(not shown), with no changes in other members of the tumour necrosis factor receptor (TNFR) superfamily or other GST isoforms. RT-PCR analysis confirmed AKNA's ability to upregulate CD40 and concomitantly downregulate GST expression (Fig. 6a). These findings were further substantiated by northern blot experiments, flow cytometry and western blotting (Fig. 6a). We did not, however,

```

MERYLSVKSLEPEAMRMEEEEEEEEEEGGDSLEVDGV 40
AATPGKAEATRVLPQCPVQAEKSHGAPLEEATEKMVSMK 80
PPGFQASLARDGHMSGLGKAEAAPGGVPPHPPGTSAA 120
SHQSSMTSLESGISERLPQKPLHRGGGPHLEETWMASPE 160
TDSGFVGSETSRVSPLTQTPEHRLSHISTAGTLAQPFAS 200
VPRDGASYPKARGSLIPRRATEPSTPRSQAQRYLSSPSGP 240
LRQRAPNFSLERTLAAEMAVPGSEFEGHKRISEQPFPNKT 280
ISPPAPAPAAAPLPCGPTETIPSFLLTRAGRDQAICELQ 320
EEVSRLRLRLSDSLHQPLQGSPTRPASAFDRPARTRGRP 360
ADSPATWGSYHSGKSTERLPGEPRGEEQIVPPGRQARRSS 400
SVPREVLRLSLSSSESELPSLPLFSEKSKTTKDSQAARDG 440
KRGVGSAGWPDRVTFRGQYTGHEYHVLRLRRSQKAMAQSP 480
VPTAGPLGPRMRVLSQGTHWDRPLIPFSVPCVVKLGLP 520
QRQMVQAQPPLGQRRPRRGEKHLQLPAPSRGASRRGRHA 560
HPPDCGIWQRPQHQLPLTSPRFPSCLIHLPLCTMRL 600
QDLPQPNQLPSGRPQPLPHQPGDTGTPSSSTWAT 633
  
```

Figure 3 AKNA has an AT-hook motif and multiple PEST domains. AKNA's deduced amino-acid sequence shows a highly hydrophilic protein with defined structural properties: a nine-amino-acid AT hook DNA-binding motif (bold) and three PEST regions (boxed) that are rich in proline (P), glutamic acid (E), serine (S) and threonine (T). The complete cDNA sequence has been deposited in the Gene Bank databases of NCBI and assigned the accession no. AF286341.

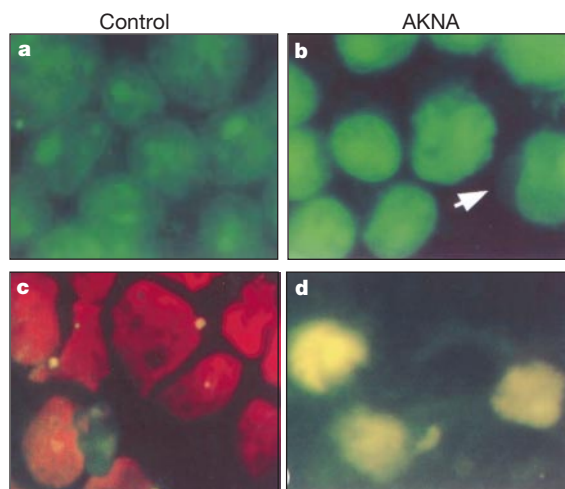


Figure 4 AKNA expression localizes in the nucleus. Fluorescence microscopy of the JY lymphoma cell line transfected with either a GFP-vector control (a) or AKNA-GFP (b). The arrow indicates the cytoplasmic shadow surrounding the bright green nucleus. c, d, Comparative red fluorescence (provided by the DNA-specific fluorochrome propidium iodide) and green fluorescence (resulting from GFP expression) analyses of GFP and AKNA-GFP transfectants.

have any evidence that AKNA could bind CD40 gene regulatory elements. As the CD40 promoter contains AT-hook binding domains²⁸, we tested the capacity of recombinant AKNA to bind to the CD40 promoter by electrophoretic mobility shift assay (EMSA). Figure 6b shows that AKNA can bind CD40 regulatory elements and supports its function as a transcription factor. Identical EMSA results were obtained with native AKNA protein extracted from tonsil B lymphocytes (Fig. 6b).

Because CD40L also contains potential A/T-rich promoter elements²⁹, is concomitantly expressed with AKNA and CD40 by germinal centre B lymphocytes and can be expressed by activated B lymphocytes^{25,26}, we hypothesized that AKNA could also bind CD40L promoter and regulate its expression. Figure 6c, d shows that AKNA also binds A/T-rich CD40L promoter elements and upregulates the expression of CD40L. These data support the role of AKNA as a key transcription factor that regulates the expression of a key receptor-ligand pair and confirm the physiological significance of AKNA during secondary immune responses. These findings are further sustained by the fact that, in addition to B lymphocytes, AKNA is expressed by CD1a⁺CD14⁻ dendritic cells, T lymphocytes and NKC (Fig. 2).

The biological significance of the short AT-hook sequence is substantiated by the demonstration that the motif is necessary and sufficient to bind DNA^{2,30}. The evidence that AKNA directly binds

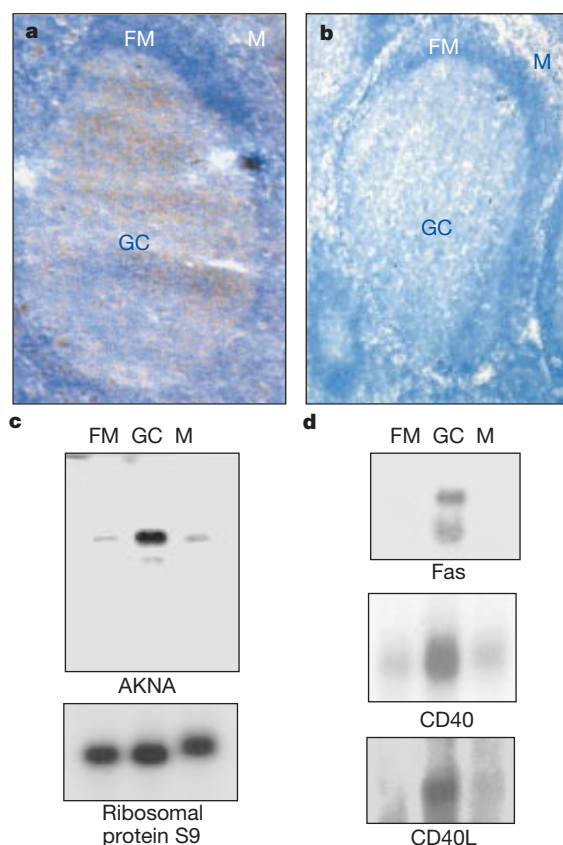


Figure 5 Concomitant expression of AKNA, Fas, CD40 and CD40L by germinal centre B lymphocytes. a, Immunohistology depicting the expression of AKNA by germinal centre (GC) cells, using anti-AKNA sera. b, Immunohistochemical analysis with control non-immune rabbit sera. c, Northern blot analysis of AKNA expression by distinct naive (which populate the follicular mantle (FM)), GC and memory (M) B-lymphocyte subsets. d, B-cell subset expression of Fas, CD40 and CD40L by northern blotting. The S9 ribosomal protein was used as an internal control for even RNA sample loading.

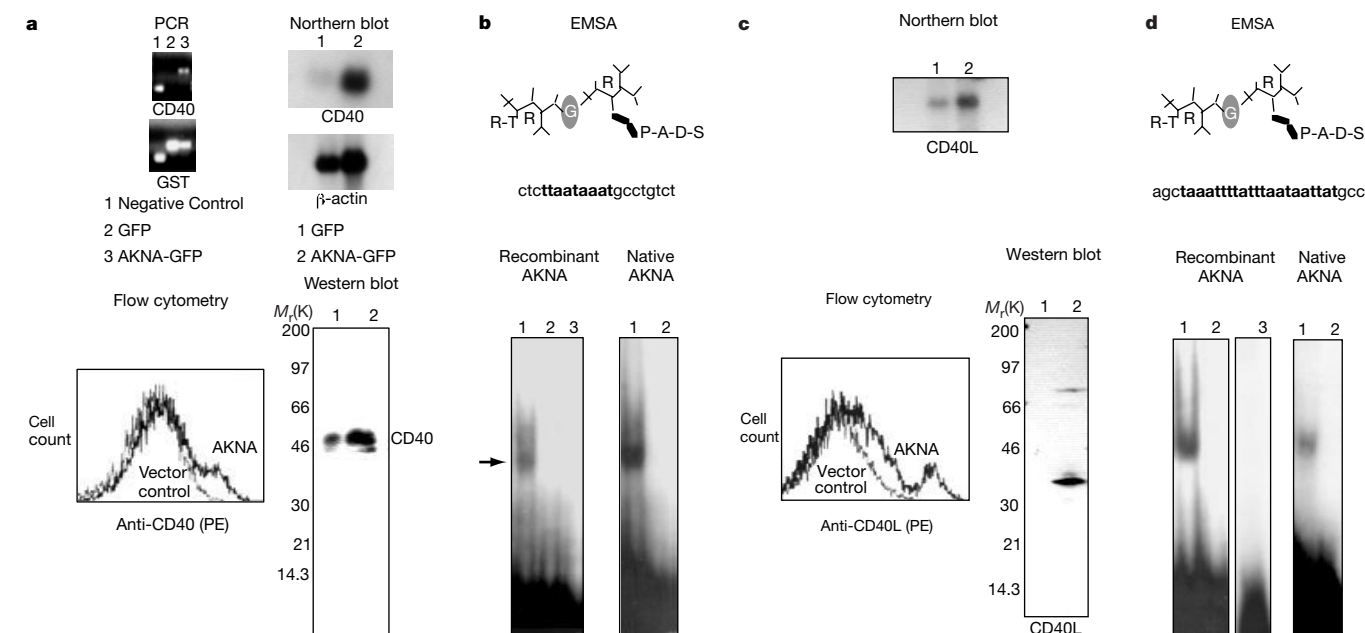


Figure 6 Upregulation of CD40 and CD40L expression. **a**, RT-PCR and northern blot analysis comparing vector control and AKNA transfectants. β -actin was used as housekeeping gene control. AKNA-mediated upregulation of CD40 was confirmed by flow cytometry and western blotting (1, GFP; 2, AKNA-GFP). Histograms of the vector control and AKNA transfectants overlap. **b**, EMSA between A/T-rich CD40 promoter and AKNA (recombinant or native proteins). 1, AKNA + 32 P CD40 A/T-rich probe; 2, competition with

the cold probe; 3, reticulocyte lysate control. Top, AKNA's AT hook (GPR-centred) motif and its binding A/T-rich oligonucleotide target. **c**, Analysis of AKNA-mediated upregulation of CD40L by flow cytometry, northern and western blotting (1, GFP; 2, AKNA-GFP). **d**, EMSA between A/T-rich CD40L regulatory element and AKNA. Top, AKNA's AT hook and CD40L promoter's A/T-rich motif. 1, AKNA + 32 P CD40L A/T-rich probe; 2, competition with the cold probe; 3, reticulocyte lysate control.

A/T-rich promoter elements strongly supports the transcription factor function of AT-hook-containing proteins. Moreover, the capacity of one transcription factor to coordinately regulate the expression of a receptor and its ligand is a new finding and provides a molecular explanation for homotypic receptor-ligand interactions. Further investigations of murine models are in progress to clarify the physiological meaning of the *in vivo* expression of AKNA. □

Methods

Cloning and sequencing

We constructed a B-lymphocyte-specific cDNA library by standard procedures using commercially obtained reagents, the Poly (A)⁺ track kit (Invitrogen), pSPORT and the Superscript-II cDNA cloning kit (Life Technologies) according to the manufacturer's specifications. The B-cell cDNA library was screened in combination with a commercially obtained random-primed T-cell cDNA library (Stratagene) using a 32 P-labelled AKNA probe generated by B-cell subset-specific RT-PCR differential display. Relevant clones were isolated, purified and automatically sequenced (DNA Sequencing Core Facility, UT-MDACC).

Northern blot analysis

Multiple tissue northern (MTN) blots (Clontech) were hybridized as recommended by the manufacturer. Alternatively, 10 μ g of total unfractionated RNA was electrophoresed, transferred and hybridized as described¹⁸ using 32 P-labelled cDNA probes. The probes for Fas, CD40 and CD40L were generated by PCR.

AKNA transfectants, and sub-cellular localization

The 1.9-kb fragment spanning the ORF of AKNA cDNA was directionally cloned into the amino-terminal region of the mammalian expression vector pEGFP-N3 (Clontech), to generate AKNA-GFP fusions. We generated GFP or GFP-AKNA transfectants by the Dotap cationic liposome gene transfer method (Boehringer Mannheim) according to the manufacturer's instructions. AKNA expression was localized by fluorescent microscopy.

Flow cytometry analysis

Vector control (GFP) and AKNA-GFP JY transfectants were stained and analysed for the expression of CD40 and CD40L. Briefly, cells were washed once in phosphate-buffered

saline (PBS) and 5% fetal bovine serum (FBS) and reacted in the dark with phycoerythrin-conjugated anti-CD40 and CD40L antibodies for 30 min at 4 °C. Stained cells were washed twice with PBS/FBS, resuspended in PBS alone and analysed by flow cytometry (Department of Immunology, UT-MDACC Core Facility).

Immunohistology

To analyse AKNA expression, we stained frozen tissue sections with affinity-purified anti-AKNA sera, followed by a secondary peroxidase-conjugated anti-rabbit reaction and development with stable DAB (Research Genetics) and Gill's haematoxylin counterstain. As a negative control, we carried out a parallel staining in which the primary antibody was an irrelevant non-immune rabbit sera.

Western blotting

Equal amounts of protein lysates, extracted from GFP or AKNA-GFP JY transfectants, were resolved by 12% SDS-polyacrylamide gel electrophoresis and electrotransferred overnight onto nitrocellulose membranes. Blots were blocked with 4% non-fat dry milk in Tris-buffered saline (TBS) with 0.2% Tween-20, for at least 1 h at 4 °C. Blots were then incubated overnight with primary antibody (CD40 or CD40L) at 4 °C, followed by three washes with TBS-Tween 20 at room temperature for 10 min each. Peroxidase-conjugated secondary antibody was added to the blots and incubated for 1 h at room temperature. Blots were washed three times and reactivity revealed by chemiluminescence (Pierce) according to the manufacturer's instructions.

Human ATLAS cDNA array

Poly (A)⁺ RNA was prepared as described above, from both GFP and GFP-AKNA stable transfectants. We generated 32 P-labelled antisense cDNA from each transfectant by a room temperature reaction using the Superscript II RT kit (Life Technologies). Atlas array analysis was performed using a kit purchased from Clontech according to the manufacturer's method.

CD40 and GST-P1 RT-PCR

Forward and reverse primers for CD40 and GST were designed and applied to amplify transcripts from GFP and AKNA-GFP transfectants. Briefly, total RNA was extracted, reverse-transcribed and amplified for 35 cycles, using RT (Life Technologies) and PCR (Perkin Elmer) kits, according to the manufacturer's instructions. CD40 oligonucleotide primers: forward 5'-AGTGGTCTCTGCCGCTGGTC-3' and reverse 5'-GTCTTGTGTTGTGCCTGCCTGTTGC-3'. GST-P1 primers: forward 5'-GAGTTTTCGCCGCCGAGTCT-3' and reverse 5'-GATCAGCAGCAAGTCCAGCAGTT-3'. Optimal annealing temperature was 60 °C for the CD40 amplification and 61 °C for GST-P1. Other parameters were as specified in the PCR kit.

EMSA

Complementary oligonucleotides (Fig. 6) with overhanging ends were synthesized, annealed and ³²P-dCTP end-label filled with Klenow DNA polymerase. Recombinant (in vitro translated) or native (extracted from freshly purified tonsil B lymphocytes) AKNA protein was incubated with poly[d(I-C)] (Boehringer-Mannheim) for 10 min at room temperature in a binding buffer containing 10 mM Tris-HCl pH 7.5, 50 mM NaCl, 5% glycerol, 5 mM MgCl₂, 0.1% Nonidet NP-40, 1 mM dithiothreitol and 50 µg of bovine serum albumin. For competition experiments, the binding reactions were incubated with 100-fold molar excess of unlabelled probe for 15 min and then the radiolabelled probe was added. The binding reaction mixes were separated in a 4% polyacrylamide gel, with a buffer containing 0.5× TBE (Tri-borate-EDTA). The gel was dried and exposed to X-ray film overnight at -70 °C.

Received 31 October 2000; accepted 5 January 2001.

1. Friedmann, M. *et al.* Organization, inducible expression, and chromosomal localization of the human HMG-1 (Y) non-histone gene. *Nucl. Acids Res.* **21**, 4259–4267 (1993).
2. Reeves, R. & Nissen, M. S. The AT-DNA-binding domain of mammalian high mobility group-1 chromosomal proteins, a novel peptide motif for recognizing DNA structure. *J. Biol. Chem.* **265**, 8573–8582 (1990).
3. Chevallier, P. PEST sequences in nuclear proteins. *Int. J. Biochem.* **25**, 479–482 (1993).
4. Liu, Y.-J., Zhang, J., Lane, P. J. L., Chan, E. Y.-T. & MacLennan, I. C. M. Sites of specific B cell activation in primary and secondary responses to T cell-dependent and T cell-independent antigens. *Eur. J. Immunol.* **21**, 2951–2962 (1991).
5. Jacob, J. & Kelsoe, G. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl) acetyl II. A common clonal origin for periarteriolar lymphoid sheath-associated foci and germinal centers. *J. Exp. Med.* **176**, 679–687 (1992).
6. MacLennan, I. C. M. Germinal centers. *Annu. Rev. Immunol.* **12**, 117–139 (1994).
7. Berek, C., Berger, A. & Apel, M. Maturation of immune responses in germinal centers. *Cell* **67**, 1121–1129 (1991).
8. McHeyzer-Williams, M. G., McLean, M. J., Lalor, P. A. & Nossal, G. J. V. Antigen-driven B cell differentiation in vivo. *J. Exp. Med.* **178**, 295–307 (1993).
9. Liu, Y. J. *et al.* Within germinal centers isotype switching of Immunoglobulin genes occurs after the onset of somatic mutation. *Immunity* **4**, 241–250 (1996).
10. Arpin, C. *et al.* Generation of memory B cells and plasma cells in vitro. *Science* **268**, 720–722 (1995).
11. Liu, Y. J. *et al.* Germinal centers contain unmutated IgD+IgM+ founder cells and hypermutated IgD+IgM- cells. *Immunity* **4**, 603–613 (1996).
12. Clark, E. A. & Ledbetter, J. A. How B and T cells talk to each other. *Nature* **367**, 425–428 (1994).
13. Grewal, I. S. & Flavell, R. A. The role of CD40 ligand in costimulation and T-cell activation. *Immunol. Rev.* **153**, 85–106 (1996).
14. Glimcher, L. & Kara, C. Sequences and factors: a guide to MHC class-II transcription. *Annu. Rev. Immunol.* **10**, 13–49 (1992).
15. Lenschow, D. J., Walunas, T. L. & Bluestone, J. A. CD28/B7 system of T cell costimulation. *Annu. Rev. Immunol.* **14**, 233–58 (1996).
16. Liang, P. & Pardee, A. B. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**, 967–971 (1992).
17. Altschul, S., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local assignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
18. Guzman-Rojas, L. *et al.* PRELI, the human homologue of the avian px19, is expressed by germinal center B lymphocytes. *Int. Immunol.* **12**, 607–612 (2000).
19. Aravind, L. & Landsman, D. AT-hook motifs identified in a wide variety of DNA-binding proteins. *Nucleic Acids Res.* **26**, 4413–4421 (1998).
20. Kolanus, W. *et al.* Alpha L beta 2 intergrin/LFA-1 binding to ICAM-1 induced by cytohesin-1, a cytoplasmic regulatory molecule. *Cell* **86**, 233–242 (1996).
21. Krammer, P. H. *et al.* The role of APO-1-mediated apoptosis in the immune system. *Immunol. Rev.* **142**, 175–191 (1994).
22. Martinez-Valdez, H. *et al.* Human germinal center B cells express the apoptosis inducing gene FAS, c-myc, p53, and Bax, but not the survival gene bcl-2. *J. Exp. Med.* **183**, 971–977 (1996).
23. Thorbecke, G. J., Amin, A. R. & Tsiaghe, V. K. Biology of germinal centers in lymphoid tissue. *FASEB J.* **8**, 832–840 (1994).
24. Craxton, A. *et al.* p38 MAPK is required for CD40-induced gene expression and proliferation in B lymphocytes. *J. Immunol.* **161**, 3225–3236 (1998).
25. Grammer, A. C., McFarland, R. D., Heaney, J., Darnell, B. F. & Lipsky, P. E. Expression, regulation, and function of B cell-expressed CD154 in germinal centers. *J. Immunol.* **163**, 4150–4159 (1999).
26. Clodi, K. *et al.* Coexpression of CD40 and CD40 ligand in B-cell lymphoma cells. *Br. J. Haematol.* **103**, 270–275 (1998).
27. Lo, H.-W. & Ali-Osman, F. Genomic cloning of hGSTP1**C*, an allelic human Pi class glutathione S-transferase gene variant and functional characterization of its retinoic acid response elements. *J. Biol. Chem.* **272**, 32743–32749 (1997).
28. Rudert, F. *et al.* Identification of a silencer, enhancer, and basal promoter region in the human CD95 (Fas/APO-1) gene. *DNA Cell Biol.* **14**, 931–937 (1995).
29. Schubert, L. A. *et al.* The human gp39 promoter. *J. Biol. Chem.* **270**, 29624–29627 (1995).
30. Huth, J. R. *et al.* The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif. *Nature Struct. Biol.* **4**, 657–665 (1997).

Acknowledgements

We thank H. N. Ananthaswamy, G. B. Mills and M. F. Wilkinson for comments and advice. This work was supported by a grant from the Leukemia and Lymphoma Society. J.C.S. and R.R. were supported by the Smith pre-doctoral fellowship. L.G.-R. and R.R. are graduate students from the Department of Immunology, School of Biological Sciences, at the Autonomous University of Nuevo Leon, Mexico. L.G.-R. is a recipient of a fellowship from the Consejo Nacional de Ciencia y Tecnologia, Mexico.

Correspondence and requests for materials should be addressed to H.M.-V. (e-mail: hmartine@mdanderson.org).

Transcription coactivator p300 binds PCNA and may have a role in DNA repair synthesis

Sameez Hasan, Paul O. Hassa, Ralph Imhof & Michael O. Hottiger

Institute of Veterinary Biochemistry, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland

The transcriptional coactivator p300 interacts with many transcription factors that participate in a broad spectrum of biological activities, such as cellular differentiation, homeostasis and growth control^{1,2}. Mouse embryos lacking both p300 alleles die around mid-gestation, with pleiotropic defects in morphogenesis, in cell differentiation and, unexpectedly, in cell proliferation because of reduced DNA synthesis³. Here we show that p300 may have a role in DNA repair synthesis through its interaction with proliferating cell nuclear antigen (PCNA). We show that *in vitro* and *in vivo* p300 forms a complex with PCNA that does not depend on the S phase of cell cycle. A large fraction of both p300 and PCNA colocalize to speckled structures in the nucleus. Furthermore, the endogenous p300–PCNA complex stimulates DNA synthesis *in vitro*. Chromatin immunoprecipitation experiments indicate that p300 is associated with freshly synthesized DNA after ultra-violet irradiation. Our results suggest that p300 may participate in chromatin remodelling at DNA lesion sites to facilitate PCNA function in DNA repair synthesis.

The inability of p300^{−/−} fibroblasts to synthesize DNA might be due to p300 being a transcriptional coactivator in the activation of S-phase-specific genes, or to p300 being an important cofactor for DNA synthesis. To examine whether p300 acts as a cofactor in DNA synthesis, we performed immunoprecipitations of p300 from HeLa nuclear extracts and analysed bound proteins by western blot using an anti-PCNA antibody. PCNA co-immunoprecipitated with the p300 but not with the control antibody (Fig. 1a). Likewise, p300 could be co-immunoprecipitated with an anti-PCNA antibody (Fig. 1b), indicating that p300 and PCNA may form a complex *in vivo*. PCNA is a homotrimer that forms a closed ring structure around duplex DNA, similar to the β-subunit of *Escherichia coli* DNA polymerase III (ref. 4; see also Supplementary Information).

To test whether the p300/PCNA interaction was S-phase-specific, cells were synchronized in G1, G1/S or late S by serum starvation or treatment with hydroxyurea, and subsequently released for 7 h. The status of cell cycle was analysed by fluorescence-activated cell-sorting analysis (data not shown). PCNA binding to p300 was detected in all tested cell-cycle phases (Fig. 1c), indicating that p300 forms a complex with PCNA that is not specifically induced for S phase. To confirm the association of PCNA and p300, we carried out colocalization studies. Antibodies against PCNA (green) and p300 (red) showed a nuclear staining of both proteins in untreated HeLa cells (Fig. 2b and c, respectively).

When confocal images were merged, subnuclear regions of strong PCNA and p300 staining coincided in yellow, speckled structures in the nucleus, indicating that both endogenous PCNA and p300 colocalize *in vivo* (Fig. 2d). This pattern of staining corresponded to that of DNA as seen in the DAPI (4',6-diamidino-2-phenylindole dihydrochloride) staining of these cells (Fig. 2a), suggesting that both proteins are in close physical proximity to DNA. No significant difference in colocalization of p300 and PCNA could be detected in S-phase-synchronized cells (data not shown). Quantification of the overlapping areas showed that not all of the PCNA and p300 molecules formed nuclear speckled structures, which indicates that PCNA and p300 are involved in other functions in the cell for which their interaction is not required.

To map the interaction domain, p300 was divided into five fragments, which we cloned and subsequently expressed as glutathione *S*-transferase (GST) fusion proteins (Fig. 3a). We carried out GST pull-down assays with these five fragments using nuclear extracts of 293 cells. Western blot analysis of bound proteins using the anti-PCNA antibody showed that fragments 4 and 5 of p300 interact with PCNA (Fig. 3b). These fragments corresponded to amino-acid residues 1,459–1,892 and 1,893–2,414 of p300. Identical experiments with bacterially expressed and purified PCNA revealed that only fragment 5, and none of the other GST p300 fragments bound to PCNA (Fig. 3c).

These results indicate that the observed binding of PCNA to fragment 4 (Fig. 3b) is probably mediated by a protein that is missing when PCNA is expressed in bacteria. Whether the binding to fragment 5 is direct or mediated by an undetectable bacterial contaminant remains to be investigated. Proteins that interact with PCNA are reported to share a PCNA-binding signature⁵. This signature, however, was not found in p300 fragment 4 or fragment 5, as is the case in other PCNA-binding proteins, such as pol δ or a subunit of human replication factor C. Further experiments revealed that the binding of PCNA to fragment 5 was stable in salt concentrations of up to 400 mM; however, stability in 400 mM salt was much less than in 200 mM salt (see Supplementary

Information).

Specific amino-acid residues of PCNA are involved in interactions with many proteins involved in DNA replication or repair⁶. We therefore examined whether specific mutations in PCNA can affect its p300 interaction. We tested four mutated isoforms of PCNA in GST pull-down experiments; the VDK188AAA, SHV43AAA and QLGI125AAAA mutants have been shown to form trimers, whereas the LAPK251AAAA mutant does not⁵. The different forms of PCNA have also been reported to vary in their ability to interact with other proteins⁵.

In pull-down experiments, p300 fragment 5 bound wild-type PCNA, VDK188AAA and LAPK251AAAA (Fig. 3d); these two mutants also interact with p21, Fen1 and DNA ligase I (ref. 5). The SHV43AAA and QLGI125AAAA mutants did not bind to GST p300 fragment 5 (Fig. 3d). Although the SHV43AAA mutant interacts with proteins such as chromatin assembly factor-1 (CAF-1), p21 and Fen1 (refs 5, 7), the QLGI125AAAA mutant does not⁵. These results confirm the specificity of the interaction between PCNA and p300, and suggest that the interaction is mediated through distinct domains on PCNA.

PCNA is an essential processivity factor for DNA polymerases (pol) δ and ϵ in eukaryotic DNA synthesis^{8,9} and is thus important for nucleotide excision repair (NER)^{10,11}. We therefore investigated whether p300-bound PCNA was functionally active and could stimulate DNA synthesis by pol δ *in vitro*. PCNA was co-immunoprecipitated from nuclear extracts using a p300-specific antibody and subjected to a PCNA-dependent pol δ DNA synthesis assay. As detected in western blot analysis (Fig. 1a), the presence of PCNA in the immunocomplex resulted in a detectable amount of PCNA-dependent DNA synthesis by pol δ (Fig. 3e, left panel), whereas immunoprecipitations using a control antibody did not. In parallel, recombinant PCNA was pulled down with different GST p300 fragments and subsequently used in the same assay. As in the

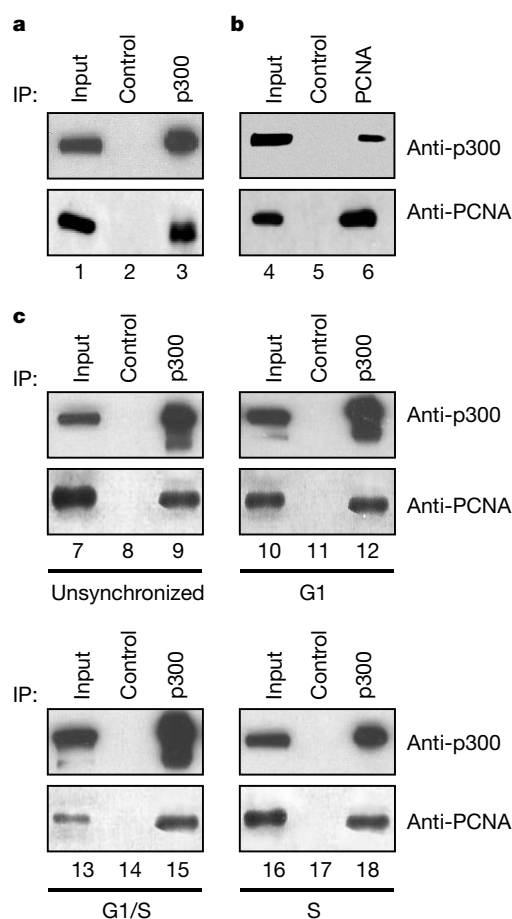


Figure 1 p300 and PCNA form a complex. **a**, p300 was immunoprecipitated from 150 μ g HeLa nuclear extracts using a p300 antibody or a control IgG. Co-immunoprecipitated proteins were visualized by western blot analysis using an anti-p300 antibody (top) or an anti-PCNA antibody (bottom). **b**, PCNA was immunoprecipitated from 150 μ g HeLa nuclear extracts using a PCNA antibody or a control IgG. **c**, HeLa cells were synchronized in G1, G1/S or S phase. Nuclear extracts were prepared and used for p300 co-immunoprecipitations.

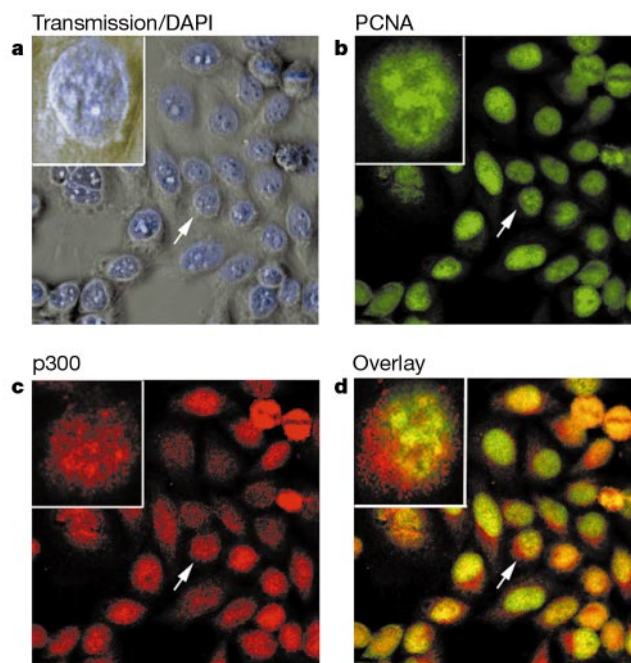


Figure 2 Colocalization of p300 and PCNA in HeLa cells *in vivo*. **a**, HeLa cells stained with DAPI (blue) for nuclei, overlaid to the phase-contrast picture. **b**, **c**, The same cells were stained with an anti-PCNA antibody (**b**) followed by a DTAF-conjugated second antibody (green), or with an anti-p300 antibody (**c**) followed by a rhodamine-coupled second antibody (red). **d**, Merge for PCNA (**b**) and p300 staining (**c**) shows colocalization of both p300 and PCNA as orange-yellow speckles in the nucleus (arrow).

western blot analysis (Fig. 3c), only fragment 5 could bind PCNA, resulting in a detectable amount of synthesized DNA after subsequent incubation of bead bound proteins with pol δ (Fig. 3e, right panel).

To show that p300 is a cofactor for DNA synthesis, we carried out chromatin immunoprecipitation (CHIP) with a p300 antibody to determine whether p300 is associated with freshly synthesized DNA *in vivo*. After fixation of the cells with formaldehyde, p300 was, as expected, detected by western blot analysis in the immunocomplex precipitated with a p300 antibody (Fig. 4a). The presence of p300-associated DNA was investigated by additionally treating the p300 immunocomplex with polynucleotide kinase and [γ - 32 P]ATP, which labelled the DNA ends of co-precipitated DNA fragments. A smear of DNA fragments with an average length of 400 base pairs (bp) was detected in the complex associated with p300 (Fig. 4b). This p300-associated radiolabelled material was sensitive to DNase I treatment, indicating that p300 is indeed associated with DNA *in vivo*

(Fig. 4b, lane 6).

To distinguish whether the p300-associated DNA was bound to p300 because of its involvement in transcription or DNA synthesis, we performed CHIP experiments with extracts of cells labelled with radioactive thymidine. We also distinguished between DNA synthesis occurring in DNA replication and DNA synthesis after DNA damage in DNA repair, by additionally irradiating a sample of cells by ultraviolet before labelling. DNA isolated from normally cycling cells showed a dTTP incorporation of 52 pmol per 10^6 cells, whereas cells treated with 25 mJ m $^{-2}$ ultraviolet radiation showed a threefold increase in thymidine incorporation (data not shown). This increase was probably due to stimulation of the cellular DNA repair machinery. In a control experiment, cells starved in G1 showed highly reduced incorporation (10% of untreated cells; data not shown). CHIPs of ultraviolet-treated cells showed a 3.5-fold increase in the amount of bound radioactively labelled DNA when immunoprecipitated with the p300 antibody in comparison

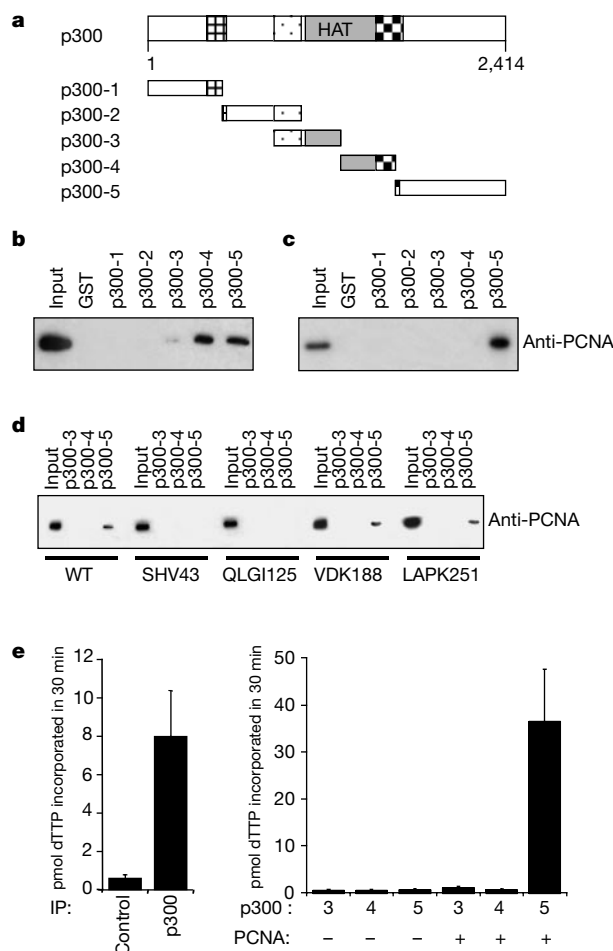


Figure 3 Mapping of the PCNA interaction domain in p300. **a**, Representation of p300. Hatched box, (cysteine/histidine)-rich region 1; dotted box, (cysteine/histidine)-rich region 2; grey box, histone acetyltransferase (HAT) domain; checked box, (cysteine/histidine)-rich region 3. **b**, GST pull-down assays using bacterially expressed GST p300 fragments 1-5 incubated with 293 nuclear extracts. Interacting PCNA was visualized by western blot analysis. **c**, Bacterially expressed PCNA (500 ng) was incubated with GST p300 fragments 1-5. Bound PCNA was visualized by western blot analysis. **d**, Pull down of wild-type (WT) and mutant PCNA by GST p300 fragments 3-5. **e**, Endogenous PCNA bound to immunoprecipitated p300 was subjected to a PCNA-dependent DNA synthesis assay with pol δ (left). Recombinant PCNA bound to GST p300 fragment 5 can be used by pol δ (right).

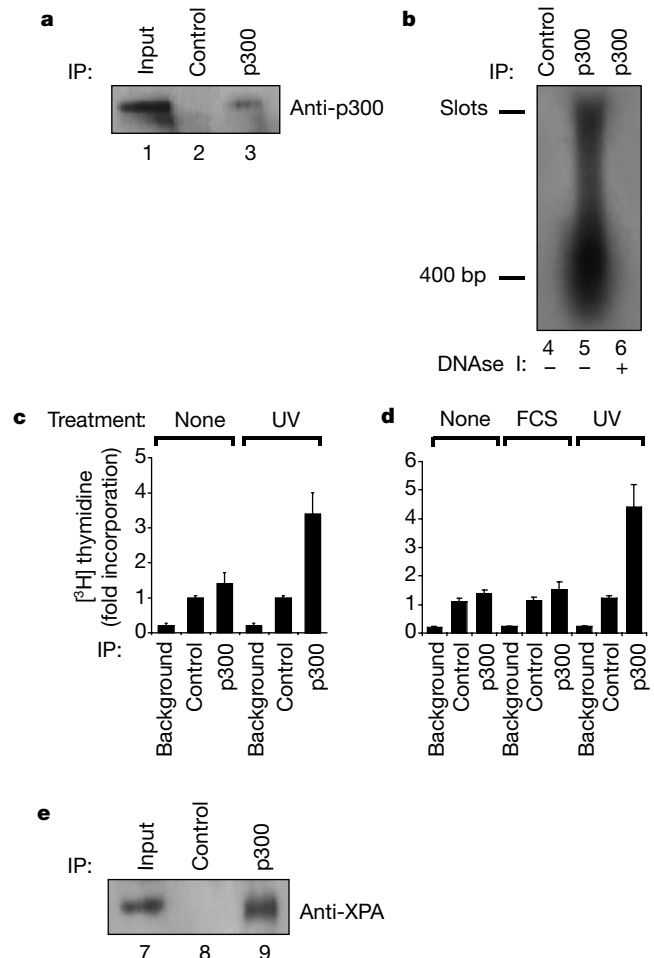


Figure 4 p300 is associated with freshly repaired DNA. **a**, Western blot analysis of p300 after immunoprecipitation of p300 using extracts of cells fixed with formaldehyde. **b**, After immunoprecipitation of p300, bound DNA was end labelled with [γ - 32 P]ATP using T4 polynucleotide kinase and visualized by autoradiography. **c**, Cells labelled with [3 H]thymidine were untreated or ultraviolet-irradiated. Subsequently, p300 was immunoprecipitated from fixed cell extracts. Incorporation of [3 H]thymidine was measured with a scintillation counter. **d**, Cells were serum starved followed by an 8-h FCS incubation or an ultraviolet pulse, parallel to labelling with [3 H]thymidine. After immunoprecipitation of p300, incorporated [3 H]thymidine was assessed. **e**, p300 and XPA form a complex. p300 was immunoprecipitated from HeLa nuclear extracts. Co-immunoprecipitated proteins were visualized by western blot analysis using an anti-XPA antibody.

with a control antibody (Fig. 4c). The experiments with the same antibodies using an extract of untreated cells showed no significant difference in the amount of immunoprecipitated labelled DNA.

The average length of the p300-bound fragments was about 400 bp, minimizing the possibility that p300 might bind to radioactive DNA indirectly at a neighbouring transcribed locus because of its function as a transcriptional coactivator. To confirm these results, we repeated experiments in which cells were arrested by fetal calf serum (FCS) depletion, and then stimulated by the addition of 20% FCS for 8 h. The addition of FCS to starved cells has been reported to activate many genes^{12–14}. Although ultraviolet treatment under these conditions stimulated the binding of p300 to labelled DNA, FCS treatment did not (Fig. 4d). These results indicate that the observed increased amount of p300-bound DNA following ultraviolet treatment is caused by the involvement of p300 in DNA repair.

To confirm this hypothesis, we carried out immunoprecipitations of p300 from HeLa nuclear extracts and analysed the bound fractions by western blot analysis for XPA, a protein known to be involved in DNA repair¹⁵. Immunoprecipitations of p300 confirmed the interaction of p300 with XPA (Fig. 4e). This interaction, like the PCNA–p300 interaction, was not dependent on ultraviolet irradiation (data not shown). Together, these results suggest that the transcriptional coactivator p300 is involved directly as a cofactor in DNA repair and that the reported association of p300 with DNA is regulated by PCNA and other factors of the cellular repair machinery. Our data do not exclude the possibility that factors involved in transcription regulation might interact simultaneously with the same p300 molecule. Whether one p300 molecule is involved in gene expression and/or DNA repair synthesis remains to be investigated.

Exposure to ultraviolet light causes damage to DNA which is processed in the NER pathway^{16,17}. In yeast, a complex of Rad7–Rad16 functions specifically in NER¹⁸. On the basis of sequence homology, Rad16 belongs to the SWI/SNF2 subfamily of DNA-dependent ATPases, a group of proteins implicated in chromatin remodelling¹⁶, with which p270 (a p300 related protein) has been reported to form a complex¹⁹. p300 also interacts with BRCA1 (ref. 20), another protein suggested to be involved in DNA repair²¹. Compacting DNA into nucleosomes and higher order structures affects the accessibility of ultraviolet-induced DNA lesions. In this respect PCNA might, after being loaded onto the damaged DNA, interact with and tether p300 to these lesions. p300, as a protein with intrinsic histone acetyltransferase activity²² and as a component of a chromatin-remodelling complex²³, might change the structure of the chromatin adjacent to DNA lesions, inducing chromatin changes that facilitate PCNA function and DNA repair synthesis, although the p300 complex is unlikely to affect DNA synthesis directly.

Proliferating cell nuclear antigen is therefore likely to be involved at several steps in DNA repair synthesis, including its association with p300 in the pre-synthesis complex, its binding to DNA polymerases^{24–26}, and the recruitment of additional components to the newly synthesized DNA to enhance chromatin formation^{7,27}. p300 might be an example of a protein that is well known for its involvement in gene activation but that also has a critical role in DNA repair. It is conceivable that the chromatin-remodelling activity associated with p300 is an important feature of the multifunctional protein in regulation of different responses to DNA damage, thus expanding its paradigm as a factor solely involved in transcription. □

Methods

Plasmids

Fragments 1–5 of p300, corresponding to amino acids 1–672, 672–1,193, 1,069–1,459, 1,459–1,892, 1,893–2,414 were amplified using polymerase chain reaction (PCR) and cloned into pGex6P2 (Clontech). The primers used were 5'-TAGCTAGCTCGAGAA

TGGCCGAGAATGTGGTGAACC-3' and 5'-ATCGATCAAGCTTTTACCTCGGATTC ATGGAACCTGGAACC-3' (GST p300-1); 5'-TA GCTAGCTCGAGGAGGGCCCTAACAT GGGACAGCCGC-3' and 5'-ATCGATCAAGCT TTTAAGTGAATAAGTGGCATCAG AGG-3' (GST p300-2); 5'-TAGCTAGCTCGAGGA GATCCAGAAATCCCTTCCTTTTCG-3' and 5'-ATCGATCAAGCTTTTACTTGGGTAT CTCTGGTCAGGAG-3' (GST p300-3); 5'-TAGCTAGCTCGAGGAAAGCCCAAGCGAC TGCAGGAATG-3' and 5'-ATCG ATCAAGCTTTTAAGCTTGAGTCTCTGGGCAAGTA GG-3' (GST p300-4); 5'-TAGCTAG CTGAGGAGCTGCTGGCCCTGTGTCCAG-3' and 5'-ATCGATCAAGCTTTTACTA GTGTATGTCTAGTGTACTCTGTG-3' (GST p300-5).

Purification of GST fusion proteins

Expression of GST fusion proteins in *E. coli* was induced by adding 0.5 mM isopropylthiogalactoside (IPTG) for 3 h to the bacterial culture in the exponential phase of growth. The bacteria were pelleted, resuspended in EBC buffer (120 mM NaCl, 50 mM Tris-HCl pH 8.0, 0.5% (v/v) NP-40 and complete EDTA-free protease inhibitor cocktail (Roche Diagnostics, Mannheim)), lysed using a French pressure cell, and pelleted to remove debris. The soluble proteins were purified by glutathione Sepharose beads and visualized by Coomassie blue staining.

In vitro protein–protein binding

We carried out GST pull-down experiments as described²⁸.

Immunoprecipitation

p300 immunoprecipitations were performed as described²⁸ with 1 µg of an anti-p300 antibody (PharMingen, 14991A; or Santa Cruz C20, sc-585) or a control immunoglobulin-γ (IgG; Sigma, M9269). For western blot analysis of bound proteins we used anti-PCNA antibody (PC-10, Santa Cruz Biotechnology) or anti-XPA antibody (sc-853, Santa Cruz Biotechnology).

Immunocytochemistry and confocal microscopy

Cells cultured on coverslips were fixed with methanol at –20 °C for 10 min followed by a brief permeabilization with 0.5% (v/v) Triton X-100 in PBS at 4 °C. The cells were washed twice with PBS, blocked for 2 h in blocking solution (2% bovine serum albumin, 0.1% (v/v) Triton X-100 in PBS) and incubated overnight with the primary antibodies at 4 °C. The samples were washed three times with blocking solution and incubated with secondary antibodies. We stained the cells with DAPI (Sigma, D9542) before mounting. For dual immunofluorescence staining, we used a p300 polyclonal antibody (C20, Santa Cruz) and a monoclonal PCNA antibody (PC-10, Santa Cruz). The secondary antibodies were anti-rabbit rhodamine-coupled (Molecular Probes, R6394) and anti-mouse dichlorotriazinyl amino fluorescein (DTAF)-coupled (Anawa, 5178–2404), respectively. Omission of the primary antibodies showed that there was minimal background staining. We analysed the preparations under an inverted laser scanning microscope (Leica).

DNA polymerase δ assay

The PCNA-dependent pol δ assay was carried out as described²⁹ using poly(dA)/oligo(dT) as a template.

[³H]thymidine uptake and chromatin immunoprecipitation

Cells were labelled with [³H]thymidine (5 µCi ml^{–1}) for at least 10 h before formaldehyde was added to the cell-culture medium at a concentration of 1% (v/v). Fixation proceeded at 22 °C for 10 min and was stopped by the addition of glycine to a concentration of 125 mM. Before collection, the cells were rinsed twice with PBS. The cells were then resuspended in solution A (10 mM HEPES pH 7.5, 10 mM EDTA, 0.5 mM EGTA and 0.25% (v/v) Triton-X) and incubated at room temperature for 10 min. Cells were taken up and incubated for 10 min at room temperature in solution B (200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 10 mM HEPES pH 7.5). The cells were lysed in 150 mM NaCl, 25 mM Tris-HCl pH 7.5, 5 mM EDTA, 1% (v/v) Triton X-100, 0.1% (w/v) SDS, 0.5% (v/v) sodium deoxycholate and complete EDTA-free protease inhibitor cocktail, and incubated on ice for 10 min, followed by sonication for 4 min in short pulses. Cell debris was removed by brief centrifugation. p300 was immunoprecipitated and the amount of synthesized DNA was measured using a scintillation counter (Kontron MR300).

Received 1 September; accepted 21 December 2000.

- Eckner, R. *et al.* Molecular cloning and functional analysis of the adenovirus E1A-associated 300-kD protein (p300) reveals a protein with properties of a transcriptional adaptor. *Genes Dev.* **8**, 869–884 (1994).
- Giordano, A. & Avantaggiati, M. L. p300 and CBP: partners for life and death. *J. Cell. Physiol.* **181**, 218–230 (1999).
- Yao, T. P. *et al.* Gene dosage-dependent embryonic development and proliferation defects in mice lacking the transcriptional integrator p300. *Cell* **93**, 361–372 (1998).
- Krishna, T. S., Kong, X. P., Gary, S., Burgers, P. M. & Kuriyan, J. Crystal structure of the eukaryotic DNA polymerase processivity factor PCNA. *Cell* **79**, 1233–1243 (1994).
- Jonsson, Z. O., Hindges, R. & Hübscher, U. Regulation of DNA replication and repair proteins through interaction with the front side of proliferating cell nuclear antigen. *EMBO J.* **17**, 2412–2425 (1998).
- Jonsson, Z. O. & Hübscher, U. Proliferating cell nuclear antigen: more than a clamp for DNA polymerases. *BioEssays* **19**, 967–975 (1997).
- Moggs, J. G. *et al.* A CAF-1-PCNA-mediated chromatin assembly pathway triggered by sensing DNA damage. *Mol. Cell. Biol.* **20**, 1206–1218 (2000).

8. Podust, V. N. & Hübscher, U. Lagging strand DNA synthesis by calf thymus DNA polymerases alpha, beta, delta and epsilon in the presence of auxiliary proteins. *Nucleic Acids Res.* **21**, 841–846 (1993).
9. Waga, S. & Stillman, B. The DNA replication fork in eukaryotic cells. *Annu. Rev. Biochem.* **67**, 721–751 (1998).
10. Nichols, A. F. & Sancar, A. Purification of PCNA as a nucleotide excision repair protein. *Nucleic Acids Res.* **20**, 2441–2446 (1992).
11. Shivji, K. K., Kenny, M. K. & Wood, R. D. Proliferating cell nuclear antigen is required for DNA excision repair. *Cell* **69**, 367–374 (1992).
12. Baldwin, A. S. Jr, Azizkhan, J. C., Jensen, D. E., Beg, A. A. & Coodly, L. R. Induction of NF-kappa B DNA-binding activity during the G0-to-G1 transition in mouse fibroblasts. *Mol. Cell. Biol.* **11**, 4943–4951 (1991).
13. Almendral, J. M. *et al.* Complexity of the early genetic response to growth factors in mouse fibroblasts. *Mol. Cell. Biol.* **8**, 2140–2148 (1988).
14. Muller, R., Bravo, R., Burckhardt, J. & Curran, T. Induction of *c-fos* gene and protein by growth factors precedes activation of *c-myc*. *Nature* **312**, 716–720 (1984).
15. Cleaver, J. E. & States, J. C. The DNA damage-recognition problem in human and other eukaryotic cells: the XPA damage binding protein. *Biochem. J.* **328**, 1–12 (1997).
16. de Laat, W. L., Jaspers, N. G. & Hoeijmakers, J. H. Molecular mechanism of nucleotide excision repair. *Genes Dev.* **13**, 768–785 (1999).
17. Sancar, A. DNA excision repair. *Annu. Rev. Biochem.* **65**, 43–81 (1996).
18. Guzder, S. N., Sung, P., Prakash, L. & Prakash, S. Yeast Rad7–Rad16 complex, specific for the nucleotide excision repair of the nontranscribed DNA strand, is an ATP-dependent DNA damage sensor. *J. Biol. Chem.* **272**, 21665–21668 (1997).
19. Dallas, P. B. *et al.* p300/CREB binding protein-related protein p270 is a component of mammalian SWI/SNF complexes. *Mol. Cell. Biol.* **18**, 3596–3603 (1998).
20. Pao, G. M., Janknecht, R., Ruffner, H., Hunter, T. & Verma, I. M. CBP/p300 interact with and function as transcriptional coactivators of BRCA1. *Proc. Natl Acad. Sci. USA* **97**, 1020–1025 (2000).
21. Zhong, Q. *et al.* Association of BRCA1 with the hRad50–hMre11–p95 complex and the DNA damage response. *Science* **285**, 747–750 (1999).
22. Ogrysko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell* **87**, 953–959 (1996).
23. Cho, H. *et al.* A human RNA polymerase II complex containing factors that modify chromatin structure. *Mol. Cell. Biol.* **18**, 5355–5363 (1998).
24. Dresler, S. L. & Frattini, M. G. DNA replication and UV-induced DNA repair synthesis in human fibroblasts are much less sensitive than DNA polymerase alpha to inhibition by butylphenyl-deoxyguanosine triphosphate. *Nucleic Acids Res.* **14**, 7093–7102 (1986).
25. Hunting, D. J., Gowans, B. J. & Dresler, S. L. DNA polymerase delta mediates excision repair in growing cells damaged with ultraviolet radiation. *Biochem. Cell Biol.* **69**, 303–308 (1991).
26. Coverley, D., Kenny, M. K., Lane, D. P. & Wood, R. D. A role for the human single-stranded DNA binding protein HSSB/RPA in an early stage of nucleotide excision repair. *Nucleic Acids Res.* **20**, 3873–3880 (1992).
27. Shibahara, K. & Stillman, B. Replication-dependent marking of DNA by PCNA facilitates CAF-1-coupled inheritance of chromatin. *Cell* **96**, 575–585 (1999).
28. Perkins, N. D. *et al.* Regulation of NF-kappaB by cyclin-dependent kinases associated with the p300 coactivator. *Science* **275**, 523–527 (1997).
29. Weiser, T. *et al.* Biochemical and functional comparison of DNA polymerases alpha, delta, and epsilon from calf thymus. *J. Biol. Chem.* **266**, 10420–10428 (1991).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank M. Stucki and Z. Jonsson for pol δ , and wild-type and mutant PCNA; U. Hübscher for critically reading the manuscript; other members of the Institute of Veterinary Biochemistry for helpful advice; and U. Ziegler, S. Koundrioukoff and M.Y. Nakano for technical help. This work was supported in part by the Swiss National Science Foundation (S.H. and P.O.H.), the UBS AG "im Auftrag eines Kunden" (R.L.) and the Olga Mayenfisch Stiftung. M.O.H. is supported by the Kanton of Zürich.

Correspondence and requests for materials should be addressed to M.O.H. (e-mail: hottiger@vetbio.unizh.ch).

Genetics and oncology

Gene expression and antibodies for diagnostics.

MACS Anti-Melanoma MicroBeads

Miltenyi Biotec www.miltenyibiotec.com

Combined cell enrichment and staining

Anti-Melanoma MicroBeads are used to enrich disseminated melanoma cells from peripheral blood, bone marrow and lymphoid tissue and for purifying melanoma cells from single-cell suspensions of skin biopsies. The antibody, clone 9.2.27, conjugated to MicroBeads, recognizes a surface antigen expressed by most melanomas. Following enrichment, melanoma cells can be analysed by immunocytochemistry, flow cytometry or RT-PCR, or taken directly into culture.

Reader Service No. 100

GeneMap

Genomic Solutions www.genomicsolutions.com

Pre-printed microarrays

The GeneMap array consists of 1,152 cancer-related genes spotted in duplicate within a 9×9, 32-patch array, with a centre-to-centre spacing between spots set at 400 µm. A control gene, the Q gene from lambda, is spotted at the A1 position of each patch, and there are eight blanks within each patch. A gene list can be downloaded from the company's website.

Reader Service No. 101

c-Met

Zymed www.zymed.com

Prognostic marker for cancers

Rabbit anti-c-Met antibody (PAD: CVD13) can be used in immunohistochemistry applications for the detection of formalin-fixed, paraffin-embedded tissues and western blotting of human samples. c-Met (HGFR) is a proto-oncogene (locus 7q21q31) and a cell-surface receptor for hepatocyte growth factor. c-Met is overexpressed in many cancers and often amplified between tumorigenesis and metastasis. As such, it may be a useful prognostic marker in breast cancer, gastric carcinoma, endometrial carcinoma and others.

Reader Service No. 102

GeneSystem320

KPL www.kpl.com

A fragmented approach

This system uses a defined set of 320 PCR primers and dedicated software available over the Internet to identify known genes or for gene discovery. A message fragmentation protocol involves isolation of actively transcribing genes, isolating a single fragment per mRNA molecule, and conducting specific combinatorial oligo PCR on the fragments. GS320 is

available for human, mouse and rat.

Reader Service No. 103

WAVE for ATM

Transgenomic www.transgenomic.co.uk

Detecting ATM exon 42 mutations

The Transgenomic WAVE Nucleic Acid Fragment Analysis System performs rapid, automated separation and quantification of single- and double-stranded nucleic-acid fragments, allowing for high throughput, greater accuracy, higher resolution, and substantial savings in cost per sample. The company's Application Note 108 covers the use of WAVE technology for the detection of the ATM (ataxia-telangiectasia) gene.

Reader Service No. 104

Beclin 1 antibody

Novus Biologicals www.novus-biologicals.com

Antibody to key autophagy protein

Novus have introduced an antibody to beclin 1. *beclin 1* is the first identified mammalian gene to mediate autophagy, and has also been attributed with tumour suppression and antiviral functions. It is speculated that Beclin 1 may work through induction of autophagy to negatively regulate tumorigenesis and control viral infections.

Reader Service No. 105

GeneXPRO contract services

InnoGenex www.innogenex.com

Gene profiling at your service

GeneXPRO will carry out assays to analyse cellular and tissue-level expression for gene expression profiling and perform target localization and validation. Gene profiling services include design, synthesis and labelling of probes, correlation of gene expression profile



Catch a WAVE.

with normal and pathogenic conditions on cells, tissues and tissue arrays, and determination of changes in gene expression in response to candidate drugs. Other services available include screening mouse monoclonal antibodies on mouse tissues, immunohistochemistry and *in situ* hybridization on human and rodent tissues.

Reader Service No. 106

COLARIS

Myriad www.myriad.com

Calculated risk

This predictive test assesses an individual's risk of developing hereditary colon cancer based on the presence of a mutation in either of two genes. The same mutations also substantially increase a woman's risk of endometrial cancer. A positive COLARIS test result indicates an 80% lifetime risk of colon cancer, and a 40–60% risk of uterine cancer in women.

Reader Service No. 107

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENTS

BLOCKERS

K⁺ Channel Blocker

E4031

Wako USA
877-714-1921

Arachidonic Acid Cascade Blocker

AA861

Wako GmbH
(49) 02131 3110

Wako www.wako-chem.co.jp

BACs • YACs • PACs cDNAs

Research Genetics provides clones and colony membranes for a number of human, mouse and rat libraries as well as a custom screening service for the libraries.

Research Genetics
U.S. or Canada 800-533-4363
U.K. 0-800-89-1393
FAX 256-536-9016

Check out our homepage at
<http://www.resgen.com>

READER ENQUIRY NO. 35

READER ENQUIRY NO. 6



Flockhart
Aiming for a
future of
bespoke drugs



Babiss
On the hunt
for genetic
fingerprints



Austin
Multitasking is
key to success in
drug design



Anderson
Assessing ethics
and recruiting
hard

Pharmacogenetics initiative galvanizes public and private sectors

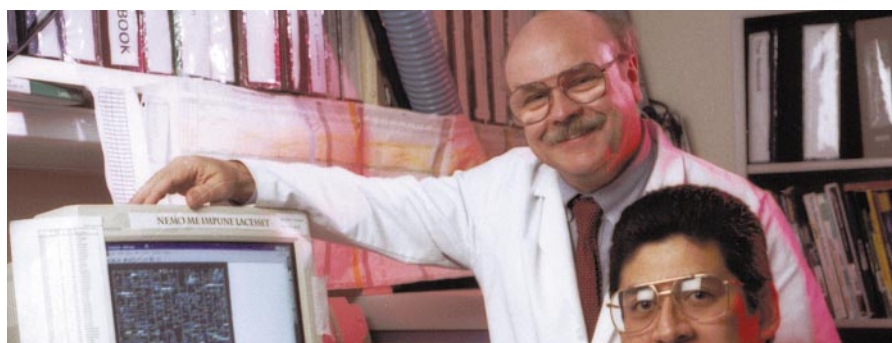
Skills in genetics and clinical research are in demand as academia and industry try to tailor treatments to their patients, says Paul Smaglik.

While serving as a Georgetown University Hospital medical resident nine years ago, David Flockhart treated a young woman who had experienced a surprise overdose — unexpected because she had followed the directions for taking her antidepressant. Investigating the drug's reaction, Flockhart and his colleagues found that the patient had a variation in a gene that codes for an enzyme essential for breaking down the drug. Because of this, her body did not produce enough of the enzyme and she could not metabolize the medicine fast enough, allowing the drug to build up to toxic levels in her system.

That experience led Flockhart into a career in pharmacogenetics. This field aims to use variations in individuals' genomes to help physicians to tailor prescriptions to each patient. Although researchers have long investigated variations in drug response, the field has become more closely linked to genomics over the past year, because of an explosion in available data.

The publication of two draft versions of the human genome has galvanized molecular biology. But the burgeoning database of single nucleotide polymorphisms (SNPs) — the small variations in genetic code that give us our individuality — may ultimately be more important to pharmacogenetics.

The possibilities of using SNPs and other variations as guides in tailoring treatments to a patient's genetic make-up have captured the interest of both public and private sectors. Cancer is considered an ideal condition on which to apply this approach, as subtle variations could make the difference between a particular dose of chemotherapy being toxic or effective. Also, seemingly similar kinds of cancers may actually have different genetic origins, warranting separate clinical approaches.



Drug tailors: David Flockhart (standing, above) is investigating reactions to drug treatments; Roche's Lee Babiss (standing, left) is searching for genetic 'fingerprints'.

The US National Institutes of Health (NIH) began a focused effort last spring when it launched the Pharmacogenetics Research Network with a series of grants to academic researchers. Cancer received the most for any specific disease, about a third of the \$53 million so far pledged over five years. The agency will make another round of grants this spring.

This year, the network expanded beyond academia to include pharmaceuticals and biotechnology. If the field is to grow further, each sector will need to expand communication, and agree on a common language of nomenclature and database standards. This could reduce redundancy — for example, by standardizing the informed-consent forms used for clinical trials. It will also attract

scientists with skills that are already in short supply.

Hybrid help

Flockhart has skills for which all three sectors are searching — he is a geneticist with both clinical and computational biology experience. He is now searching for similar individuals to round out his research team. He received \$1.3 million last May to discern which breast cancer patients best respond to the drug tamoxifen.

Right now, Flockhart suspects that there is more research devoted to narrowing down such pools of potential SNPs using computational biology than there is looking at which SNPs actually play roles in clinical trials. He recommends that scientists interested in pharmacogenetics learn to live with

careers and recruitment

one foot in basic science and the other in applied research. "You've got to be able to see the clinic and the DNA," he says. "That's where the work is needed."

Finding relationships like the one between reactions to antidepressants and to chemotherapy will require computing strategies beyond those that bioinformatics currently offers, says Russ Altman, associate professor of medicine and computer science at Stanford University. If the same metabolizing enzyme breaks down an antidepressant and a chemotherapeutic drug, a deficiency in that enzyme could cause the chemotherapeutic agent to be unnecessarily toxic. Altman received \$1.6 million from the NIH last May to build the database that will house the initiative's information.

He wants to design the database so that it will automatically check for clues to how data from projects on areas as disparate as cancer and asthma are related. Doing so will require an artificial intelligence expert, software engineers and database managers.

Mark Ratain, chair of the NIH's pharmacogenetics research network and University of Chicago professor of medicine, has less stringent requirements for his trial — but equal difficulty in meeting them. His clinical investigation of several chemotherapeutic agents needs a constant supply of postdocs with molecular biology experience who are willing to learn about clinical research. "We can't find people," Ratain says.

He suspects that scientists have a hard time straddling sectors — a necessity for pharmacogenetics research. "You have to keep a balance between what's scientifically interesting and therefore of academic value, and what is of interest to a commercial entity such as pharma or biotech," he says. Ratain suspects that more pharmacogenetics specialists are being attracted to pharma than to academia.

Pharma quest

But the difficulty in finding people with the right skills is not limited to academia. "I'm actually engaged in finding people to do this, and it's tough," says Christopher Austin, director of genomic pharmacology at Merck Research Laboratories in West Point, Pennsylvania.

Austin, a member of the NIH initiative's industry liaison group, notes that, although many PhDs trained in molecular biology techniques are searching for jobs, few have a detailed knowledge of methodologies in population genetics. And only a small subset of those also have clinical backgrounds.

Both recruitment and scientific issues in pharmacogenetics will become more difficult as scientists turn their attention from toxicity to efficacy, because efficacy probably has more components — both in terms of the genetics behind the disease and the drug's interactions with multiple biochemical

pathways. "If there is this level of clinical complexity in terms of drug response, then that means that you need somebody who understands the clinical stuff at least reasonably well," says Austin. "That's where we are having trouble finding people."

As pharmacogenetics moves into understanding diseases in which multiple genes play roles, biostatisticians will become increasingly in demand, says Penelope Manasco, a research director at Glaxo SmithKline. In oncology, most genetic characterization is performed on the tumour. Increasingly, scientists will focus on the patient's genome, which will require more complex science to interpret the tangled interactions.

Don Anderson, director of pharmacogenomics and medical genomics at Pharmacia in Kalamazoo, Michigan, suspects that new careers in pharmacogenetics will extend beyond traditional lab-bound ones, although the company is also recruiting clinical specialists, statistical geneticists and bioinformaticians to support its pharmacogenetics programme. For example, the company is looking for marketing people with scientific backgrounds who can explain the benefits of the pharmacogenetics approach to customers.

It is also investing in bioethics specialists because standards vary from country to country. Sorting out the bioethics of pharmacogenetics is important, because clinical trials now only allow scientists to collect information about specific research questions. Some scientists think more broad collection of data would benefit the field, because such data will reveal unexpected links between different diseases and drugs. But ethicists would need to address consent and privacy issues to allow more open-ended data collection.

Testing instrumental

Not all pharma companies are rushing to bring in new people to bolster their clinical pharmacogenetics approach. Hoffmann-La Roche is building its own database of drug toxicity, but is doing so without additional recruiting, says Lee Babiss, Roche vice-president of preclinical research and development. The company is taking 200 known toxic drugs, profiling them in a variety of rat models and human tissue culture, then generating a gene expression pattern Babiss calls a "fingerprint". That will help to weed out drugs that would potentially be toxic to people with similar genetic fingerprints.

Pharmacogenetics needs better diagnostics to move the use of such markers from experimental settings into clinics, according to Babiss. Roche is moving in that direction.



Merck's Christopher Austin finds it hard to get people with the right skills for pharmacogenetics.

"We're developing an awful lot of potential diagnostics in the area of tumour profiling and moving away from the concept of characterizing a tumour based on its origin," Babiss says.

Many biotech companies want to develop tests, as the market for genetic testing in doctors' surgeries is potentially huge. But for the most part, there are no fast, cheap, easy tests available. For now, many biotech companies are carrying out testing for other companies.

Variagenics, a biotech company based in Cambridge, Massachusetts, performs tests for clinical research organizations. The company focuses on compounds whose molecular pharmacology is already well known, including a number of chemotherapeutic agents. That knowledge allows the company to concentrate on the unknown — the genetic variations that interact with the medicine, says Vincent Stanton, Variagenics president and chief executive. Variagenics is the sole biotech company represented on the pharmacogenetic research network's industry liaison group.

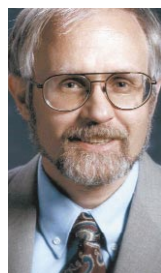
There are now three ways to look at genetic variability: examining polymorphisms, including SNPs; looking at changes in gene copy (both loss and gain); and profiling gene expression using chips. "There is no one technology that can capture all three of those things," says Stanton.

Variagenics runs all three kinds of experiments. For example, to determine whether a patient has hereditary non-polyposis colorectal cancer, which does not respond to chemotherapy in the same way as more common forms, the company looks for repeated segments of DNA in the cancerous cells that do not appear in healthy ones.

Developing and validating such tests will fuel the company's short-term growth, says Stanton. He expects the company to double its number of employees over the next year. Getting such tests into more and larger clinical trials could well determine whether or not pharmacogenetics is viable scientifically. And developing effective tests for doctors' clinics may determine whether the approach is profitable commercially. ■

Paul Smaglik is editor of *Naturejobs*.

► <http://www.nigms.nih.gov/pharmacogenetics/index.html>



Pharmacia's Don Anderson is investing in bioethics.